## G4017 Syllabus

**Course title**: **Deep Sequencing**

Offered in Fall 2017
Credit: 3
Time: Tuesdays, 2pm - 5pm.
Office hours: Friday 2:30-4pm

**Schedule**: Each week, a two-hour lecture and a one-hour journal club discussion; project presentations in the last week.

**Instructors**:
Dr. Yufeng Shen, ys2411@cumc.columbia.edu, ICRC, 803B
Dr. Peter Sims, pas2182@columbia.edu, Lasker Building, 2nd Floor, Room 203AC
Dr. Chaolin Zhang, cz2294@columbia.edu, P&S, 4th Floor, Room 4-448

TA: Alexander Hsieh (alh2194)

**Overview**:

Next-generation sequencing (NGS) has become ubiquitous in biomedical research with numerous applications. This course will provide an in-depth introduction to principles of modern sequencing, key computational algorithms and statistical models, and applications in disease genetics, cancer and fundamental biology. It will cover genome, exome, transcriptome, and epigenome sequencing approaches. Emphasis will be placed on understanding the interplay between experimental design, data acquisition, and data analysis so that students can apply these powerful tools in their own research.

*Topics*:

- History and development of modern sequencing technologies.
- Introduction to statistics and algorithms.
- Genome and exome sequencing and genetics of human diseases, such as autism, birth defects, and cancer.
- RNA-Seq and gene expression regulation (analysis and laboratory techniques for expression profiling, CLIP-Seq, ribosome profiling, micro-RNA profiling).
- Emerging single molecule sequencing technologies and single cell analysis.

*Goals*:

- To acquire introductory knowledge of modern high-throughput sequencing approaches including instrumentation and laboratory techniques.
- To understand computational and statistical methods for analyzing genome sequencing data.
- To formulate a biological question and investigate it by analyzing existing sequencing data.

*Requirements*:

This course is intended for graduate students or senior undergraduates interested in learning state-of-the-art sequencing approaches and their applications in biological and medical research.

**Required**: Basic knowledge in molecular biology, probability, and statistics
**Preferred**: working knowledge of UNIX and a programming language (R, MATLAB, Python, Ruby, or Perl). Basic knowledge of genetics would also be helpful.

**Grading**: Class participation: 10%; Journal club: 20%; Mid-term exam: 20%; Final project and presentation: 50%

Lecture - 1 09/05
**Introduction to High-Throughput Sequencing Technologies** (Sims).

Reading: (1) Sanger et al., DNA sequencing with chain-terminating inhibitors. PNAS, 74, 5463-5467 (1977). (2) Rothberg et al. Genome sequencing in microfabricated high-density picoliter reactors. Nature, 437, 376-380 (2005). (3) Shendure et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science, 309, 1728-1732 (2005). (4) Bentley et al., Accurate whole genome sequencing using reversible terminator chemistry. Nature, 456, 53-59 (2008). (5) Brock et al., Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature, 487, 190-195 (2012). (6) Laszlo et al., Decoding long nanopore sequencing reads of natural DNA. Nature Biotechnology, 32, 829-833 (2014). (7) Loose, M., Malla, S., Stout, M. Real-time selective sequencing using nanopore technology. Nature Methods, 13, 751 (2016).

JOURNAL CLUB SIGNUP

Lecture - 2 09/12
**Experimental Techniques for Genomics** (Sims). POLL for bootcamp 1.

Reading: (1) Venter et al., The sequence of the human genome. Science, 291, 1304-1351 (2001). (2) Wheeler et al., The complete genome of an individual by massively parallel DNA sequencing. Nature, 452, 872-876 (2008). (3) Porreca et al., Multiplex amplification of large sets of human exons. Nature Methods, 4, 931-936 (2007). (4) Mamanova et al., Target-enrichment strategies for next-generation sequencing. Nature Methods, 7, 111-118 (2010). (5) Adey et al., Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biology, 11, R119 (2010).

Journal Club: Amini et al. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. Nature Genetics, 46, 1343-1349 (2014).

Lecture - 3 09/19

**Technologies for Transcriptomics and Regulatory Genomics** (Sims).

Reading: (1) Mortazavi et al., Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods, 5, 621-628 (2008). (2) Licatalosi et al., HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature, 456, 464-469 (2008). (3) Ingolia et al., Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science, 324, 218-223 (2009). (4) Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. Science 316, 1497-1502 (2007). (5) Rhee, H.S., Pugh, B.F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell, 147, 1408-1419 (2011). (6) Crawford, G.E. et al., Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Research, 16, 123-131 (2006). (7) Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods, 10, 1213-1218 (2013). (8) Frommer M. et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. PNAS, 89, 1827-1831 (1992).

Journal Club: Mills E.W., Wangen J., Green R., Ingolia N.T.. Dynamic regulation of a ribosome rescue pathway in erythroid cells and platelets. Cell Reports, 17, 1-10 (2016).

09/20 Need change to Day time. Make a poll
**Optional Bootcamp on UNIX and R** (Sims/Zhang/Shen)

Lecture - 4 09/26
**Project Ideas, Team Forming, Computational Bootcamp** (Sims/Zhang/Shen).

Reading: Web-based resources will be available.

Teams of 3-4 students with balanced expertise will formed for Final Projects. Each team will propose research topic.

Instructors will conduct an optional computational bootcamp to provide basic instruction in using the UNIX operating system, scripting, and analyzing a simple data set with R.

Lecture - 5 10/03
**Technologies for Regulatory Transcriptomics** (Zhang).

Reading: (1) Heiman, M. et al. A translational profiling approach for the molecular characterization of CNS cell types. (2) Licatalosi et al., HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature, 456, 464-469 (2008). Cell 135, 738-748 (2008). (3) Mayer, A. et al. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell 161, 541-54 (2015). (4) Derti, A. et al. A quantitative atlas of polyadenylation in five mammals. Genome Res 22, 1173-1183 (2012). (5) Subtelny, A.O., Eichhorn, S.W., Chen, G.R., Sive, H. & Bartel, D.P. Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature 508, 66-71 (2014). (6) Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J.S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature 505, 701 (2014). (7) Lu, Z. et al. RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. Cell 165, 1267-79 (2016).

Journal Club: 1. Y. Diao et al., A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nat Methods 14, 629 (2017). 2. S. Zhu et al., Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. Nat Biotechnol 34, 1279 (2016).

Lecture - 6 10/10

**Statistical methods for Analyzing Sequencing Data** (Shen).

Reading: (1) All short articles in Points of Significance (http://www.nature.com/collections/qghhqm/pointsofsignificance)
(2) Eddy, S.R. What is a hidden Markov model? Nature Biotechnology, PMID: 15470472 (2004). (3) Krzywinski, M., Altman, N. Power and sample size. Nature Methods (2013). (4) Noble, W.S. How does multiple testing correction work? Nature Biotechnology, PMID: 20010596 (2009). (5) (optional) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, by Durbin, Eddy et al.

Journal Club: 1) Krzywinski, M., Altman, N. Power and sample size. Nature Methods (2013). 2) Noble, W.S. How does multiple testing correction work? Nature Biotechnology, PMID: 20010596 (2009).

Lecture - 7 10/17

**Statistical Models for Analyzing Sequencing Data (continued)** (Zhang).

Reading: (1) Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100, 9440 - 9445 (2003). (2) Bilmes, J.A. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. (Berkley, CA: International Computer Science Institute., 1998).

(3) Marioni, J., Mason, C., Mane, S., Stephens, M. & Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 18, 1509-1517 (2008). (4) Martin, E.R. et al. SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. Bioinformatics 26, 2803-2810 (2010). (5) Maaten, L.v.d. Visualizing Data using t-SNE. J Mach Learn Res, 2579-2605 (2008). (6) (optional) Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987-93 (2011). (6) (optional) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, by Durbin, Eddy et al.

Journal club: A. H. Rizvi et al., Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. Nat Biotechnol 35, 551 (2017).

Lecture - 8 10/24
**Genome Sequencing and Algorithms** (Shen).
Following a one hour lecture, each Final Project team will submit a written abstract and present a five-minute summary of their plans. A take-home midterm exam will be distributed. The midterm exam must be completed and returned at the beginning of class in Week 9 (10/31 at 2:00 pm).

Reading: (1) A framework for variation discovery and genotyping using next-generation DNA sequencing data, by DePristo et al 2011, Nature Genetics, doi: 10.1038/ng.806. PMID: 21478889. (2) Fast and accurate short read alignment with Burrows-Wheeler transform, by Li and Durbin, 2009, Bioinformatics, doi: 10.1093/bioinformatics/btp324. PMID: 19451168. (3) Exome sequencing identifies the cause of a mendelian disorder, by Ng et al 2010, Nature Genetics, doi: 10.1038/ng.499. PMID: 19915526. (4) Poplin et al (2016) Creating a universal SNP and small indel variant caller with deep neural networks, *bioRxiv* (5) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples, by Cibulskis et al 2013, Nature Biotech, doi: 10.1038/nbt.2514. PMID: 23396013. (6) A map of human genome variation from population-scale sequencing, by Abecasis et al 2010, Nature, doi: 10.1038/nature09534. PMID: 20981092.

Journal Club: Poplin et al (2016) Creating a universal SNP and small indel variant caller with deep neural networks, *bioRxiv*

Lecture - 9 10/31
**Genome Sequencing and Applications in Genetics Studies of Human Diseases** (Shen).

Take-home midterm exam is due at the beginning of class (2:00 pm).

**Reading**: (1) A framework for the interpretation of de novo mutation in human disease, by Samocha et al 2014, Nature Genetics, doi:10.1038/ng.3050 (2) Analysis of protein-coding genetic variation in 60,706 humans, by Lek et al 2016, Nature, doi:10.1038/nature19057 (3) Synaptic, transcriptional and chromatin genes disrupted in autism, by De Rubeis et al 2014, Nature, doi: 10.1038/nature13772. PMID: 25363760. (4) Prevalence and architecture of de novo mutations in developmental disorders, by The DDD Study, 2017, Nature, doi: 10.1038/nature21062. (5) The contribution of de novo coding mutations to autism spectrum disorder, by Iossifov et al, 2014, Nature, doi: 10.1038/nature13908. (6) Excess of rare, inherited truncating mutations in autism, by Krumm et al, 2015, Nature Genetics, doi: 10.1038/ng.3303.

Journal Club: Analysis of protein-coding genetic variation in 60,706 humans, by Lek et al 2016, Nature, doi:10.1038/nature19057

Lecture - 10 11/7
**Genome Sequencing and Applications in Genetics Studies of Human Diseases** (Shen).

Reading: (1) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome, by Ley et al, 2008, Nature, doi: 10.1038/nature07485. PMID: 18987736. (2) The life history of 21 breast cancers, by Nik-Zainal et al 2012, Cell, doi: 10.1016/j.cell.2012.04.023. (3) Signatures of mutational processes in human cancer, by Alexandrov et al, 2013, Nature, doi:10.1038/nature12477; . (4) Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome, by Davoli et al (2013), Cell, PMID: 24183448.

Journal Club: Signatures of mutational processes in human cancer, by Alexandrov et al, 2013, Nature, doi:10.1038/nature12477

Lecture - 11 11/14

**Transcriptome Sequencing: Algorithms and Applications in Studying Regulation of Gene Expression** (Zhang).

Reading: (1) Wu, J., Anczukow, O., Krainer, A.R., Zhang, M.Q. & Zhang, C. OLego: Fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. Nucleic Acids Res 41, 5149-5163 (2013). (2) Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21 (2013). (3) Anders, S. et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nature Protocols 8, 1765-1786 (2013). (4) Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34, 525-7 (2016). (5) Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotech 28, 511-515 (2010). (6)

Hiller, D., Jiang, H., Xu, W., Wong, W.H. Identifiability of isoform deconvolution from junction arrays and RNA-Seq. Bioinformatics, 25, 3056-3059 (2009).

Journal Club: N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34, 525 (2016).

Lecture - 12 11/21

**Transcriptome Sequencing: Algorithms and Applications in Studying Regulation of Gene Expression** (Zhang).

Reading: (1) Shen, S. et al. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A 111, E5593-601 (2014). (2) Zhang, C. & Darnell, R.B. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat Biotech 29, 607-614 (2011). (3) Widespread RNA and DNA sequence difference in the human transcriptome. Science, 333, 53-58 (2011) (4) Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506-11 (2013). (5) GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348, 648-60 (2015). (6) Li, Y.I. et al. RNA splicing is a primary link between genetic variation and disease. Science 352, 600-604 (2016).

Journal Club: A. C. McMahon et al., TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. Cell 165, 742 (2016).

Lecture - 13 11/28
**Emerging Technology for Sequencing and Single Cell Analysis** (Sims).

Reading: (1) Tang et al., mRNA-Seq whole- transcriptome analysis of a single cell. Nature Methods, 6, 377-382 (2009). (2) Pollen et al., Low- coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nature Biotechnology, 32, 1053-1058 (2014). (3) Zong et al., Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science, 338, 1622-1626 (2012). (4) Mocosko et al., Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell, 161, 1202-1214 (2015). (5) Cusanovich et al., Multiplex single cell profiling of chromatin accessibility by combinatorial molecular indexing. Science, 348, 910-914 (2015).

Journal Club: Deng, Q., Ramskold, D., Reinius, B., Sandberg, R. Single-cell RNA-Seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science, 343, 193-196 (2014).

Lecture - 14  12/5
**Final Project Presentations** (Shen/Zhang/Sims).

Collect feedbacks