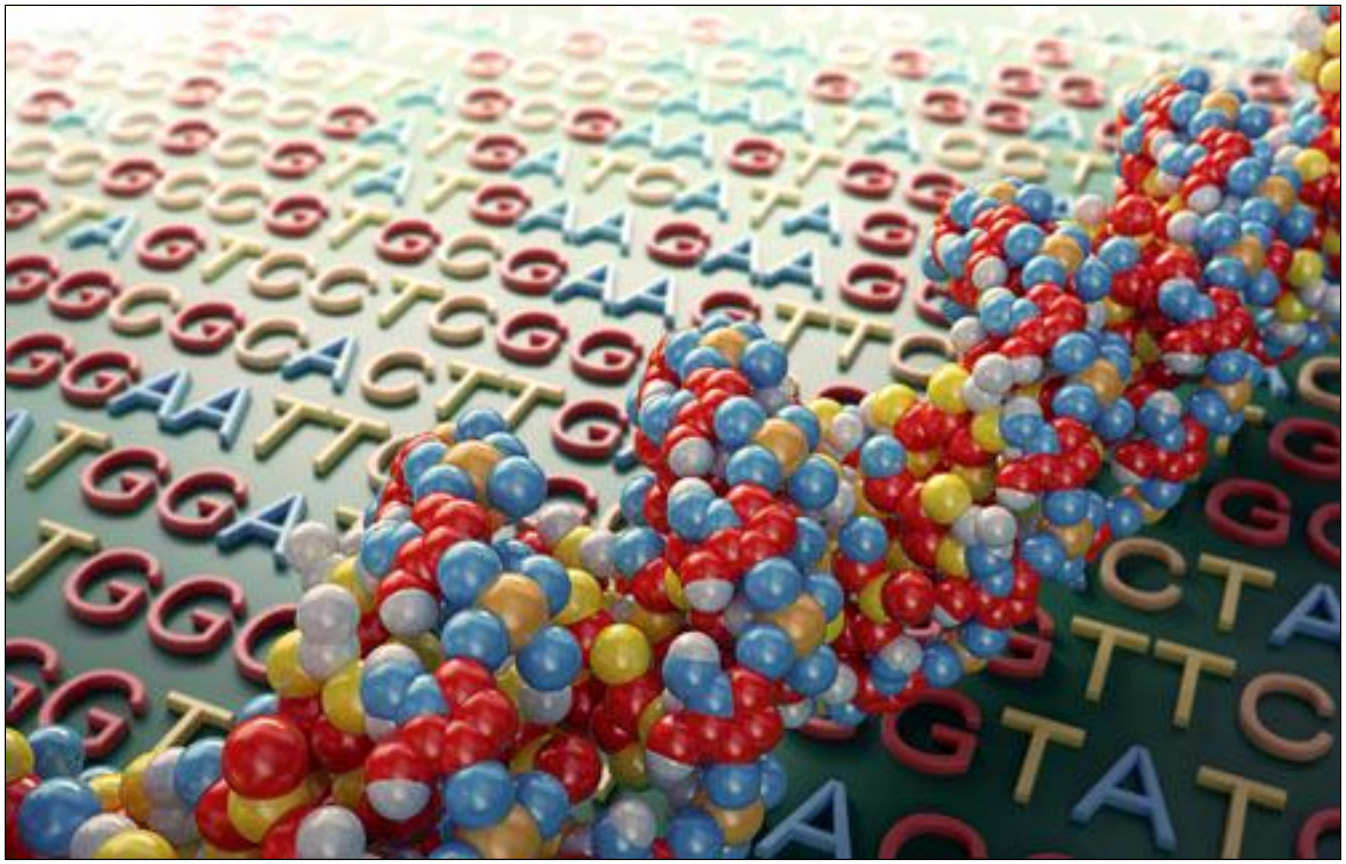
Satellite Education Workshop (SW1):

Application of NGS Technologies for Whole Transcriptome and Genome Analysis



Saturday March 2, 2013
Palm Springs, CA

Since the company was founded in 2000, NuGEN has been a leader in the development of novel genomic sample preparation solutions that address the challenges of working with clinical samples to advance discovery and drive scientific breakthroughs.

Through innovative product design, NuGEN's reagent systems allow more relevant data to be generated in less time, with less sample input. Samples of varying input amounts and quality from a variety of sources can be accessed for analysis on all leading genomic platforms.

Next-Gen Sequencing

Advances in NGS are being leveraged by researchers to address a diverse range of biological questions about the genome, transcriptome and epigenome. Access to precious clinical samples plays an important role in these applications. NuGEN's portfolio of NGS products are designed to enable generation of usable data from low input or challenging clinical samples, such as laser capture micro-dissected and formalin-fixed paraffin-embedded (FFPE) tissues. NuGEN provides technology that encompasses a range of NGS applications including RNA-Seq, whole genome DNA sequencing, target capture, ChIP-Seq, and Methyl-Seq.

Microarrray and qPCR

Microarray and qPCR technologies have a number of useful applications in the field of life science. Expression profiling has been utilized to functionally characterize biological systems in basic research and discover biomarkers for disease and treatment management in clinical settings, while copy number analysis has been used to elucidate genomic changes implicated in human disease, drug resistance and developmental defects. NuGEN's robust sample preparation solutions for RNA and DNA enable the analysis of specimens that are limited, or of poor quality on all leading microarray and qPCR analysis platforms.

Automation

Laboratory automation liberates scientists from performing routine tasks and expands research capabilities through improved reproducibility, and the ability to analyze more samples. The Mondrian SP+ Workstation, perfectly sized for any laboratory, provides benchtop microfluidics for NGS sample preparation through a simple load and go user interface. For the demand of high throughput applications, NuGEN offers customized scripts for a range of sample types on third party platforms. These methods are developed to streamline sample preparation technologies for Next Gen Sequencing, microarray analysis and qPCR. Automation size reagent kits are available in both catalog and custom formats to support customer needs.

rubicongenomics.com

RUBICON GENOMICS
breaking the sample barrier

Rubicon ThruPLEX™-FD

Single Tube Sample Preparation for Illumina® NGS Platforms
For Nanogram to Picogram Inputs of Fragmented dsDNA

Single tube, simple protocol is faster than other NGS preps like TruSeq™

VALUE PROPOSITION

Faster time to results

Less idle time

More results per day

Improved cost per result

Sample Applications

- Chromatin immunoprecipitates (ChIP)
- Mechanically sheared DNA
- Enzymatically fragmented DNA
- Cell free DNA in plasma and other biofluids
- FFPE DNA
- Flow-sorted chromosomes and cells
- cDNA
- SureSelect Enrichment

Contact Us

sales@rubicongenomics.com

rubicongenomics.com

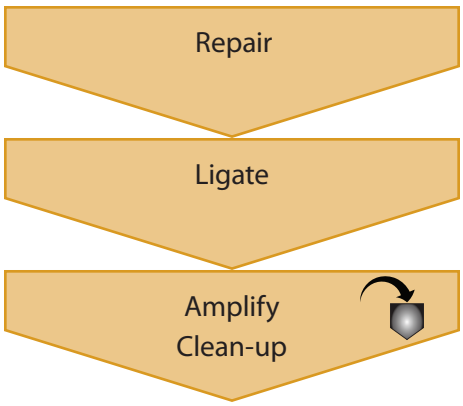
RUBICON GENOMICS
4355 VARSITY DRIVE, SUITE E
ANN ARBOR, MICHIGAN 48108

P +1.734.677.4845
F +1.734.477.9902

ThruPLEX-FD™ improves laboratory efficiency

- Picogram input levels (100x more sensitive)
- Less hands-on time required
- 1 person, 192 samples, 1 day
- Easily automatable with indexing capabilities

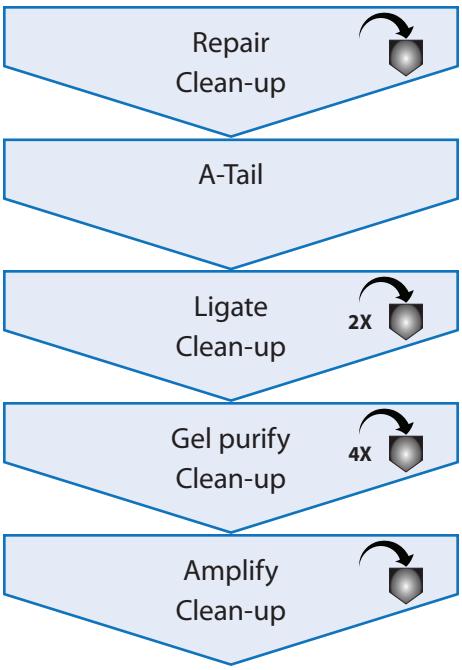
ThruPLEX-FD



NGS-READY DNA
< 3 hours
1 tube

50 pg – 50 ng DNA input recommended
1 user-supplied reagent
22 pipetting steps
No sample transfers and no columns
No Intermediate purifications
1 clean-up at end
192 sample throughput per day per tech

Illumina TruSeq™ DNA LS Prep Kit*



NGS-READY DNA
11 hours

1 µg DNA input recommended
13 user-supplied reagents
> 50 pipetting steps
8 sample transfers
5 purifications
1 clean-up at end

* Workflow and stats from TruSeq™ DNA Prep Guide 1502686 C, August 2012

Workshop Description:

This satellite workshop is intended to discuss current and future trends in the NGS platforms for whole transcriptome and genome analysis. The workshop will educate those new to the area and scientists with experience to build upon their current knowledge on the rapidly evolving NGS technology. The morning portion of the workshop will provide an overview of the available sequencing technologies and their applications in whole genome, transcriptome, and mitochondrial genome analysis. The afternoon portion of the workshop will consist of two concurrent breakout sessions. Session one will provide firsthand information on the performance of the currently available sample processing solutions for preparing sequencing libraries with limited input material and high throughput tools. Session two is dedicated to presentations on bioinformatic programs for sequencing data analysis with speakers demonstrating the use of platforms such as Galaxy and MG-RAST.

Organizer:

Nalini Raghavachari , Ph.D,
DNA Sequencing and Genomics Core Facility, NHLBI, NIH

Education Committee Liaison:

David Needleman, Ph. D,
IBR-Genetic Analysis Core Facility, Eastern Regional Research Center, USDA

Insert Half page ad for Illumina here – provided separately

Instructors:

Steve Scherer, PhD, *Baylor College of Medicine*

Steve Scherer was born and raised in Colorado. He graduated from Colorado College with a B.A. in Biology and then worked in the oil business for six years as an engineer in various Middle Eastern countries while based in Europe. He earned his Ph.D. in molecular biology from the Biochemistry Department at Baylor College of Medicine (BCM) under Dr. C. Thomas Caskey. After a short postdoctoral fellowship at BCM and another at the National Institutes of Health, Dr. Richard Gibbs, Director of the Human Genome Sequencing Center, recruited Dr. Scherer back to BCM in 1997. Dr. Scherer has served in numerous roles and currently manages multiple projects in pharmacogenomics, cancer, hearing loss and clinical application of sequencing technologies

Lee Jun Wong, PhD, *Baylor College of Medicine*

Dr. Wong received her BS in Biochemistry from National Taiwan University and Ph.D. in Chemistry from the Ohio State University, Columbus, Ohio, USA. She did her postdoctoral training in biochemical sciences at Fox Chase Cancer Center, Philadelphia, with the Nobel Laureate Dr. Irwin Rose, and Princeton University with Dr. Bruce Alberts, the past president of US National Academy of Science, followed by 14 years of tenure at the University of Massachusetts. She then re-directed her career to Medical Genetics and received her training at Baylor College of Medicine, Houston, Texas. She is certified by the American Board of Medical Genetics in the specialties of Clinical Molecular Genetics and Biochemical Genetics. She has been the Director of the Molecular Diagnostic Laboratory at Childrens Hospital Los Angeles (Univ Southern California) and Georgetown University (Washington DC) for 10 years before rejoining Baylor College of Medicine as a tenured full professor at the Department of Molecular and Human Genetics in 2005 to head the Mitochondrial/Metabolic Molecular Diagnostic Laboratory. Her research interest is in the area of mitochondrial genetics and function in disease, aging, and cancer. Her latest research development is the study of the molecular pathogenetic mechanism of mitochondrial DNA depletion syndrome, the functional significance of altered mitochondria from cancer cells, and the association of mitochondrial genetic background with the risk of diseases and cancers. She developed the MitoMet oligonucleotide array targeted to coding regions of genes related to metabolic and mitochondrial diseases for simultaneous detection of copy number changes in both nuclear and mitochondrial genomes. She has also developed the one-step comprehensive diagnosis of mitochondrial disorders by target gene enrichment followed by Next Generation massive parallel sequencing. More recently, she established CLIA/CAP validated next generation sequencing technology for clinical diagnoses of several groups of metabolic disorders including mitochondrial, glycogen storage, cholestasis, metabolic myopathy, mtDNA depletion and maintenance of mtDNA integrity, Mitome and other metabolic pathways.

Steve Kain, PhD – *Nugen Technologies*

Dr. Steve Kain is the Director of Product Management for NuGEN Technologies, a company focused on the development of sample preparation solutions for genomics research. In this role he manages all product launches and technical marketing for next-generation sequencing as well as microarray applications. Prior to NuGEN, Steve held product management and R&D roles with Complete Genomics, Agilent, and Clontech Laboratories. Steve has a Ph.D. in Biochemistry from UC Riverside and MBA from Santa Clara University.

John Langmore, PhD – *Rubicon Genomics*

Dr. John P. Langmore PhD co-founded Rubicon Genomics, Inc., in 1998 and serves as its Chief Scientific Officer. Dr. Langmore is a Co-inventor of the core technology of Rubicon Genomics and is responsible for the development of technical and commercial partnerships. Dr. Langmore served as Vice President of Commercial Development Rubicon Genomics, Inc. Prior to forming Rubicon, Dr. Langmore was a Professor in the Department of Biology and Research Scientist in the Biophysics ... Research Division at the University of Michigan, Ann Arbor. Dr. Langmore's research was in biochemical and microscopy studies of DNA and DNA-protein complexes while at the Medical Research Council Laboratory of Molecular Biology in Cambridge, England, and at the University of Michigan. During his tenure as Chair of the Biophysics Research Division at the University of Michigan, he oversaw dramatic expansion of the physical space, faculty, and graduate program. He has served as a member of National Science Foundation, American Cancer Society, and NIH peer review panels, and has over 40 publications. Dr. Langmore received his Ph.D. from the University of Chicago.

Ali Mortazavi, PhD – *University of California, Irvine*

Ali Mortazavi is an Asst. professor in the department of Developmental & Cell Biology in the University of California – Irvine. His expertise is in the applications of genomics, computation, and sequencing technologies for the analysis of transcriptional regulation in developmental biology.

Uma Dandrekar, PhD - *University of Los Angeles, David Geffen School of Medicine*

Uma Dandekar is the Asst. Director Genoseq Core from 2006 in the department of Human Genetics in the University of Los Angeles, David Geffen School of Medicine. The core provides services to research groups on the UCLA campus and in the broader scientific community throughout the world. The GenoSeq Core is involved in a broad range of scientific research including human as well as non human organisms. The core is equipped with Sanger Sequencers, Roche 454 and Illumina’s Miseq NGS platforms as well as Fluidigm’s Biomark instrument. The core is also equipped with ancillary equipment like Taqman, Pyrosequencer, Bioanalyzer.

Rebecca Laborde, PhD – *Mayo Clinic*

Rebecca Laborde completed a 4 year postdoctoral research fellowship in the department of Otorhinolaryngology at Mayo Clinic in Rochester MN. Her research focused the applications of Next Generation sequencing technologies to address questions related to risk factor exposure in the clinical setting. She is currently a principle research scientist in the department of hematology at Mayo Clinic. Her work is now focused on molecular studies of myeloid malignancies. Rebecca also holds the position of lecturer at the University of Minnesota in Rochester MN.

Dave Clements, PhD – *Emory University*

Dave Clements (Emory University, USA) coordinates training and outreach efforts for the Galaxy Project. He has led training and outreach activities for two large open source projects (Galaxy and GMOD) and has worked in genomics, gene expression, anatomy ontologies, genome visualization, and biological databases.

Folker Meyer, PhD – *Argonne National Labs*

Folker Meyer is a computational biologist at Argonne National Laboratory and a senior fellow at the Computation Institute at the University of Chicago. He was trained as a computer scientists and with that came his interest in building software systems. He now is interested in building systems that further our understanding of biological data sets. In the past he has been best known for his leadership role in the development of the GenDB genome annotation system and the design and implementation of Bielefeld University's high-performance computing facility. Currently he is most interested in comparative analysis of large numbers of microbial genomes.

Shawn O'Neil, PhD – *Oregon State University*

A Michigan native, Shawn O'Neil received his B.S. in Computer Science from Northern Michigan University in 2005. He received his M.S. in Computer Science and Engineering at the University of Notre Dame in 2009 studying algorithmic risk management for inventory control, and his Ph.D. in Computer Science specializing in Bioinformatics from the University of Notre Dame in 2012 studying transcriptomics, the effects of climate change on natural populations, and algorithms for haplotyping and related problems. He is currently a faculty research assistant in the Center for Genome Research and Biocomputing at Oregon State University, where he serves as a bioinformatics trainer, instructor, and researcher

Agenda

Morning Session

7:00am - 8:00am	**Continental Breakfast**
8:00am – 8:15am	**Introduction** Nalini Raghavachari, *NIH*
8:15am – 9:15am	**Recent Advances in Second and Third Generation Sequencing** Steve Scherer, *Human Genome Sequencing Center, Baylor College of Medicine*
9:15am – 10:00am	**Comprehensive Analysis of Mitochondrial Genome** Lee Jun Wong, *Baylor College of Medicine*
10:00am - 10:30am	**AM BREAK**
10:30am – 11:00am	**Integrated NGS Sample Preparation Solutions for Limiting Amounts of RNA and DNA** Steve Kain, *Nugen Technologies* – Giga Sponsor
11:00am – 11:30am	**Challenges and Solutions for Sequencing DNA and RNA from Clinical Samples** John Langmore, *Rubicon Genomics* – Giga Sponsor
11:30am – 12:00pm	**Panel Discussion**
12:00pm - 1:00 pm	**Lunch**

Afternoon Breakout Session 1

1:00pm – 2:00pm	**Application of NGS in HIV/Forensics** Steve Scherer, *Baylor College of Medicine*
2:00pm – 3:00pm	**Dynamics of Open Chromatin Accessibility During Myeloid Differentiation** Ali Mortazavi, *University of California, Irvine*
3:00pm - 3:30pm	**Afternoon Break**
3:30pm - 4:15pm	**Targeted Sequencing Removing Sample Prep Bottle Neck Using Access Arrays** Uma Dandrekar, *University of California* (sponsored by Fluidigm)
4:15pm – 5:00pm	**Clinical Application of NGS in Dual Genome Analysis** Lee Jun Wong, *Baylor College of Medicine*

Afternoon Breakout Session 2

1:00pm – 2:00pm	**Breaking the Data Analysis Bottleneck: Solutions that Work for RNA and Exome Sequencing** Rebecca Laborde, *Mayo Clinic* (Sponsored by Perkin Elmer)
2:00pm – 3:00pm	**Accessible, Transparent, and Reproducible analysis with Galaxy** Dave Clements, *Emory University*
3:00pm - 3:30pm	**Afternoon Break**
3:30pm - 4:15pm	**An In-depth Look on Metagenome Analysis with MG-RAST** Folker Meyer, *Argonne National Labs*
4:15pm – 5:00pm	**Assessing De-Novo Transcriptome Assembly Metrics** Shawn O'Neil, *Oregon State University*

ABSTRACTS

Recent Advances in Second and Third Generation Sequencing

Steve Scherer, Harsha Doddapaneni, Kim Worley, Stephen Richards, Jeffrey Rogers, Fuli Yu, Min Wang, Jeffery Reid, Donna Muzny, David Wheeler, Michael Metzker, Eric Boerwinkle and Richard Gibbs

From the earliest phases of the Human Genome Project (HGP), the Baylor College of Medicine's Human Genome Sequencing Center (BCM-HGSC) has taken a lead in bringing new technologies to the challenges posed by high-throughput nucleotide sequencing. We have pioneered the use of alternative fluorophores, were early adopters of next generation sequencers (NGS) and the first to sequence an individual human. More recently, we have spearheaded application of third generation technologies and successfully translated NGS to the clinical setting. This application of multiple sequencing platforms will continue to provide insight into fields such as cancer genetics, the human microbiome and pharmacogenomics as sequencing technologies enable personalized medicine.

Sequencing Applications in HIV Forensics

Steve Scherer, the BCM-HGSC, Donna Muzny, Eric Boerwinkle, Richard Gibbs and Michael Metzker

In that individuals infected with HIV-1 harbor a dynamically evolving population of related viral genomes, classical DNA profile matching is of limited use. Instead, phylogenetic methods have been used to determine the viral pattern of descent to determine the relatedness of samples. Over more than a decade, we have used first, Sanger dideoxynucleotide sequencing and more recently, next generation sequencing (NGS) technologies to resolve high profile cases involving HIV transmission. The implications and development of a forensics pathogen toolkit will be discussed.

Next Generation Sequencing Analysis of the Complex Dual Genome Mitochondrial Disorders: Technical Approach

Lee-Jun C. Wong, *Baylor College of Medicine*

Mitochondrial disorders are by far the most genetically heterogeneous group of diseases, involving two genomes, the 16.6 kb mitochondrial genome and ~1,500 genes encoded in the nuclear genome. For maternally inherited mitochondrial DNA disorders, a complete molecular diagnosis requires several different methods for the detection and quantification of mtDNA point mutations and large deletions. For mitochondrial disorders caused by autosomal recessive , dominant, and X-linked nuclear genes, the diagnosis has relied on clinical, biochemical, and molecular studies to point to a group of candidate genes followed by stepwise Sanger sequencing of the candidate genes

one-by-one. The development of Next Generation Sequencing (NGS) has revolutionized the diagnostic approach. Using massively parallel sequencing (MPS) analysis of the entire mitochondrial genome, mtDNA point mutations and deletions can be detected and quantified in one single step. The NGS approach also allows simultaneous analyses of a group of genes or the whole exome, thus, the mutations in causative gene(s) can be identified in one-step. New approaches make genetic analyses much faster and more efficient. Huge amounts of sequencing data produced by the new technologies brought new challenges to bioinformatics, analytical pipelines, and interpretation of numerous novel variants. This lecture will focus on the technical approach to the molecular diagnosis of mitochondrial disorders

Next Generation Sequencing Analysis of the Complex Dual Genome Mitochondrial Disorders: Clinical Application

Lee-Jun C. Wong, *Baylor College of Medicine*

Mitochondrial disorders are by far the most genetically heterogeneous group of diseases, involving two genomes, the 16.6 kb mitochondrial genome and ~1,500 genes encoded in the nuclear genome. For maternally inherited mitochondrial DNA disorders, a complete molecular diagnosis requires several different methods for the detection and quantification of mtDNA point mutations and large deletions. For mitochondrial disorders caused by autosomal recessive , dominant, and X-linked nuclear genes, the diagnosis has relied on clinical, biochemical, and molecular studies to point to a group of candidate genes followed by stepwise Sanger sequencing of the candidate genes one-by-one. The development of Next Generation Sequencing (NGS) has revolutionized the diagnostic approach. Using massively parallel sequencing (MPS) analysis of the entire mitochondrial genome, mtDNA point mutations and deletions can be detected and quantified in one single step. The NGS approach also allows simultaneous analyses of a group of genes or the whole exome, thus, the mutations in causative gene(s) can be identified in one-step. New approaches make genetic analyses much faster and more efficient. Huge amounts of sequencing data produced by the new technologies brought new challenges to bioinformatics, analytical pipelines, and interpretation of numerous novel variants. This lecture will focus on the clinical application of NGS to the molecular diagnosis of mitochondrial disorders.

Accessible, Transparent, and Reproducible Analysis with Galaxy

Dave Clements, *Emory University*

Galaxy is a freely available analysis and data integration framework for accessible, reproducible, and transparent biomedical research. Galaxy helps researchers extract insight from the copious amounts of data now being generated in biology, and it does this without requiring users to learn computer programming or command line interfaces.

Galaxy also supports integration of data from multiple data sources, and analysis sharing and reproducibility. Galaxy is available as a free public web site and as an open-source software package that can be installed locally or on a compute cloud.

In this workshop participants will use Galaxy to perform complex bioinformatics analyses of next generation sequencing (NGS) data, and to experiment with multiple options and parameter settings. The workshop will also demonstrate Galaxy's visual analytics capabilities, showing how to use Galaxy visualization to iteratively guide analysis.

Integrated NGS Sample Preparation Solutions for Limiting Amounts of RNA and DNA

Steve Kain, *Nugen Technologies*

Genomics sample preparation solutions that are amenable to low input workflows are essential for the analysis of limiting biological samples down to the single cell level. This presentation will highlight the application of several core technologies developed by NuGEN to enable low input studies for both transcriptome analysis as well as DNA sequencing applications.

Challenges and Solutions for Sequencing DNA and RNA From Clinical Samples

John Langmore, *Rubicon Genomics*

Next generation sequencing (NGS) enables analysis of any or all parts of the human genome in a way that opens new opportunities for clinical research and testing that is at least more comprehensive and accurate, and possibly faster and less expensive than previously possible. However NGS requires microgram quantities of molecular libraries comprised of short genomic DNA flanked by adaptor sequences using different forms of ligation-mediated PCR. For clinical DNA samples, which are limited to nanogram amounts of fragmented DNA of variable quantity and quality, preparation of high-quality libraries is difficult due to the inefficiencies and biases of conventional repair, ligation, and amplification enzymology. Original library preparation methods developed for research applications are too slow and require excessive amounts of intact input DNA to be useful in clinical applications. New generations of library preparation methods have addressed the time issue, but have not yet proven to give sufficiently uniform coverage with low noise to be successful with clinical samples. For example, the quantity of DNA in maternal plasma is typically 1 – 10 ng/mL, of which only 2 – 10% is of fetal origin; and the effective quantity of amplifiable DNA in a formalin-fixed surgical section is typically 10 – 150 ng, of which only 5 – 50% is tumor derived. This presentation explains practical challenges to clinical NGS, different methods of synthesizing and amplifying libraries, and simple metrics by which NGS library method can be validated for many clinical applications—addressing concerns about preparation time, simplicity, sensitivity, coverage, and robustness as well as real-world variables in library synthesis and amplification.

Practical workflow metrics for NGS library preparations are: 1) elapsed time; 2) simplicity (number of pipetting steps and purifications); and 3) daily throughput. Practical NGS library performance metrics are 1) density of high-quality, unique, genomic reads; 2) background from non-genomic DNA; 3) library diversity; 4) uniformity of coverage at low resolution; 5) uniformity of coverage across the range of GC composition; 6) sensitivity of sequencing metrics to variations in the amount and degradation of gDNA; and 7) reagent lot-to-lot variation in library yield, background, and sequence quality. Some current library preparation methods achieve high scores on these metrics. Fortunately the performance metrics are easily measured using qPCR during library synthesis, and highly-multiplexed NGS with <1M reads per sample. They are useful not only for evaluating and optimizing different library preparation methods, but also for quality control of individual patient samples.

Different methods of preparing NGS libraries from DNA of plasma, FFPE tissue and single cells have been evaluated using these metrics. Specific types of libraries produced in 1 – 3 hours elapsed time with fewer than 4 steps can yield reproducible whole genome and targeted sequencing results from less than 10 ng of human gDNA. For clinical applications, off-the-shelf library preparation kits and protocols are capable of producing high-quality results over a hundred-fold range of gDNA input, capable of pre-qualifying up to 96 clinical samples in a single MiSeq flow cell, and capable of detecting copy number variations, point mutations, and translocations in single cells, plasma, FFPE, and other samples.

Targeted Sequencing- Removing Sample Prep Bottleneck Using Access Array

Uma Dandrekar, *University of Los Angeles, David Geffen School of Medicine*

Next generation sequencing (NGS) has gained tremendous popularity in the past few years.NGS is a high throughput platform which requires large number of amplicons and samples to be pooled together for higher efficiency. Manually generating large number of amplified products becomes a rate limiting factor. In our laboratory we have compared the use of manual processing, liquid handling robotic instruments to the Fluidigm Access Array.

A review of the techniques revealed that although the automated liquid handler can be used for generating the amplicons required for downstream NGS processing, it is not an economical. The Fluidigm Access Array provides an efficient and economical solution to address cost and efficiency issues. In the past 2 years the use of this instrument has significantly cut down our total costs. This includes higher sample throughput, faster turnaround time, bar-coding, saving in reagent and consumable costs and hands on time. The additional services provided to support the Access Array e.g. primer design, optimization and synthesis, wet testing have made it easy to plan and execute new projects in a timely fashion. In summary the access array provides a robust, cost effective and efficient solution for amplicon/library preparation for NGS projects**.**

Dynamics of Open Chromatin Accessibility During Myeloid Differentiation

Ali Mortazavi, *University of California, Irvine*

The dynamics of vertebrate developmental commitments are ultimately translated within the cell into the remodeling of chromatin accessibility of cis-regulatory modules (CRMs) across the genome. As progenitor cells commit to differentiation, these CRMs turn off one set of genes and turn on another set. Deep DNase-seq identifies hypersensitive regions that are the hallmarks of active CRMs and can be mined footprints of DNA-binding elements. To better understand the dynamics of cell-fate commitment and differentiation, we are using DNase-seq in several time courses of differentiation of mouse C2C12 cells from myoblasts to myocytes to understand which CRMs are activated, repressed and to track changes in protein-DNA footprints as the cells irreversibly differentiate. Combined with RNA-seq and ChIP-seq of key sequence-specific transcription factors such as MyoD and Myogenin, we are mapping and analyzing the dynamics of the underlying gene-regulatory network at the level of the enhancers and promoters that underlie myogenesis.

Assessing De-Novo Transcriptome Assembly Metrics

Shawn O'Neil, *Oregon State University*

Transcriptome sequencing represents a great resource for the study of non-model species, and many metrics have been used to evaluate and compare de-novo transcriptome assemblies. Unfortunately, it is still unclear which of these metrics accurately reflect assembly quality. We address this question by simulating transcriptome sequencing and assembling the reads using both a "perfect" assembler and a modern transcriptome assembler. Given these comparative assemblies, we evaluate a variety of metrics to determine whether they 1) reveal perfect assemblies to be of higher quality, and 2) reveal perfect assemblies to be of higher quality as data quantity increases. We find that several commonly used metrics are not consistent with these expectations. We also develop and evaluate a number of novel metrics, and explore how they can speak to assembly quality in unique ways. These results provide an important review of transcriptome assembly metrics and give researchers the information needed to produce high quality reference datasets.

Breaking the Data Analysis Bottleneck: Solutions that Work for RNA and Exome Sequencing

Rebecca Laborde, *Mayo Clinic* (Sponsored by Perkin Elmer)

The treatment of oropharyngeal cancer is complicated by the significant differences in prognosis related to risk factor exposure, including tobacco and alcohol use and

infection with human papillomavirus. Patients with a limited exposure to traditional risk factors of smoking and drinking that are also positive for HPV infections have an increased rate of treatment response and 5 year survival. In an effort to investigate both the molecular mechanism of risk factor exposure and develop better tools for risk factor patient stratification, we employed Next Generation Sequencing technologies. Data generated using mRNA-Seq was combined with Exome sequencing data and allowed us to examine patterns of gene expression related to tobacco exposure, HPV infection and the presence of damaging variants. This data demonstrated differing patterns of gene expression and variant detection in patient groups representing current and never smokers. The most interesting finding was that former smoking patient groups, with a minimum of 10 year cessation history, displayed gene expression patterns more similar to never smokers. This finding suggests that gene expression patterns revert to a non-smoking phenotype following smoking cessation and provides an immediate clinical tool for encouraging patient participation in cessation programs.

Metagenome Analysis with MG-RAST

Folker Meyer, *Argonne National Laboratory*
A Wilke, J Bischof, N Desai, M D'Souza, E Glass, K Handley, T Harrison, A Howe, K Keegan, H Matthews, T Paczian, W Tan, W Trimble, J Wilkening

The availability of inexpensive next generation sequencing has enabled unprecedented quantities of environmental sequencing. Unlike previously, the bottleneck for most experiments is the computational analysis of sequence data. The MG-RAST server has been used by hundreds of groups to analyze their metagenomic data. Suitable for amplicon metagenomic data sets (16s and 18s), whole metagenome shotgun and metatranscriptomic data sets, MG-RAST has analyzed over 60,000 data sets equaling approximately 20 Tbp. The field of metagenomics is transforming our ability to study the enormous biomass and diversity of microbial life around us. Understanding this microbial world will lead to advances and practical applications in a broad range of fields. Metagenomic sequencing, provides unprecedented access to the thousands (or even millions) of microbes in an environment. Unlike 16S SSU rRNA amplicon sequencing, metagenomic sequencing (whole shotgun sequencing) provides information on not only who is in a community but what they are doing, extending understanding of community structure towards interactions within an environment. This talk will discuss the MG-RAST analysis pipeline starting from quality control assessment to annotation and an overview the interactive tools for comparative analysis. MG-RAST has analyzed over 60,000 WGS and amplicon datasets equaling approximately 20 Tbp

Acknowledgements

We would like to thank our GIGA sponsors *Nugen Technologies* and *Rubicon Genomics*, as well as our MEGA Sponsor *Illumina*, for their generous support of the workshop.

We also thank *Fluidigm* and *Perkin Elmer* for sponsoring the travel for a workshop speaker. We would like to extend our gratitude to the Education Committee and the ABRF for providing critical and necessary support for this workshop.

GIGA Sponsors

MEGA Sponsor