

GENOME TECHNOLOGY

Genomic Regulation Technical Guide

**A TROUBLESHOOTING GUIDE:
EXPERTS SHARE THEIR ADVICE ON GENOMIC
REGULATION RESEARCH FROM EPIGENETIC
MODIFICATIONS TO MICRORNAS.**



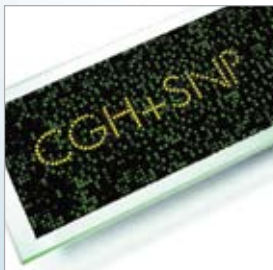
M E T H O D S

&

A

VARIANTS BETTER DEFINED

INTRODUCING AGILENT'S COMPLETE CYTOGENETICS RESEARCH SOLUTION:



- Detect more chromosomal aberrations on a single microarray, with Agilent's SurePrint G3 CGH+SNP Microarray
- Catalog microarray content developed with ISCA or add your own custom CGH content
- Software designed to enable streamlined analysis, external database comparisons, and customizable reports

Learn more at www.agilent.com/genomics/cgh_snp

© Agilent Technologies, Inc. 2011 This item is not approved for use in diagnostic procedures. User is responsible for obtaining regulatory approval or clearance from the appropriate authorities prior to diagnostic use.



Table of Contents

Letter from the Editor.....	5
Index of Experts.....	5
Q1 Which histone modification-mapping techniques do you use, and why?	7
Q2 Which genome-scale methylation mapping techniques do you use, and why?	7
Q3 Which method do you use to identify and validate microRNA targets?	8
Q4 What measures do you take to ensure reproducibility in your functional analyses of genomic regulation? ..	8
Q5 When using high-throughput sequencing, how do you balance coverage versus cost for any given experiment?	9
Q6 What are your protocols for data storage and sharing?	10
Genomic Regulation Grants	11
Resources	12

Attention Academics!

**Your access to
Clinical Sequencing News
is now free thanks to underwriting
funds from Knome.**

For the latest news on the clinical strategies of sequencing vendors as well as the adoption of sequencing technology by established clinical labs, check out <http://www.genomeweb.com/newsletter/clinical-sequencing-news>



Cambridge Healthtech Institute's Sixth Annual

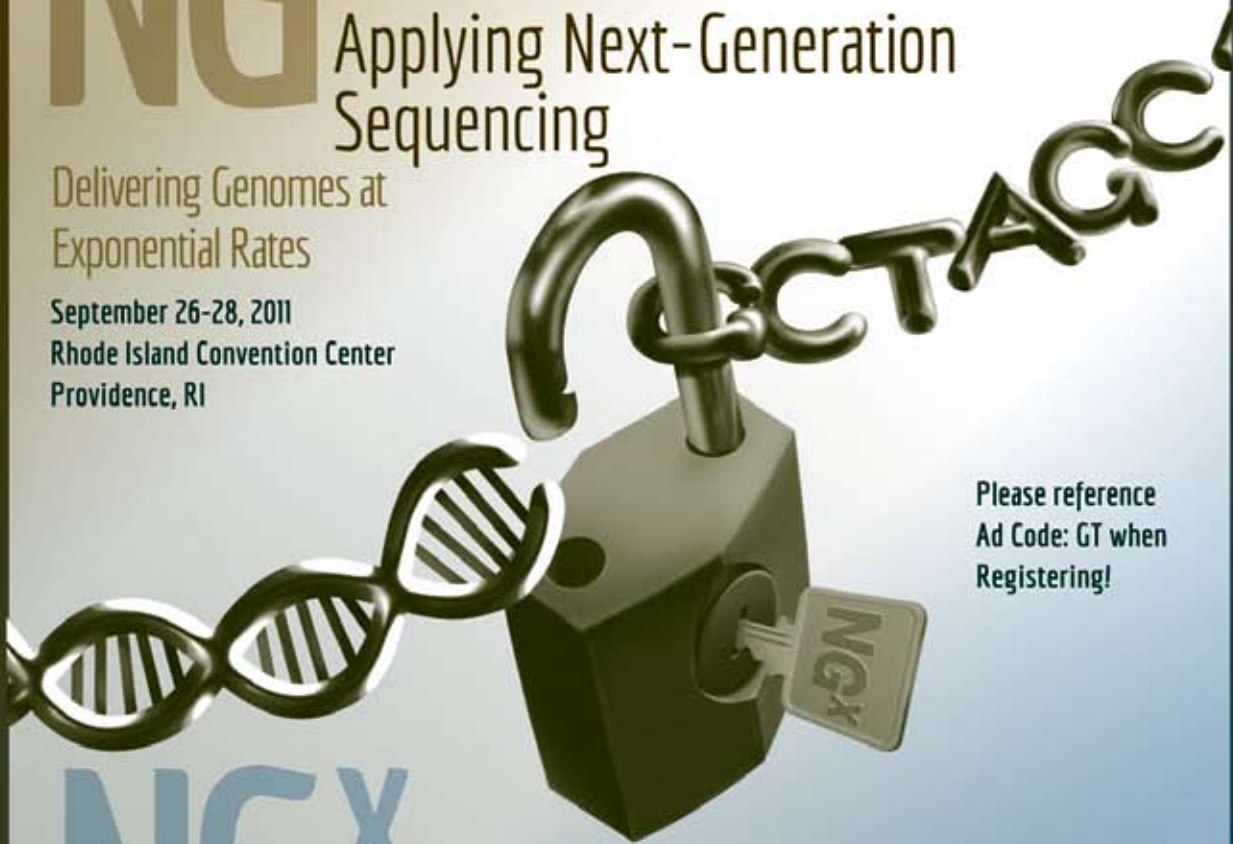
NGX

Applying Next-Generation Sequencing

Delivering Genomes at Exponential Rates

September 26-28, 2011
Rhode Island Convention Center
Providence, RI

Register by August 26
and Save up to **\$300**



Please reference
Ad Code: GT when
Registering!

NGX

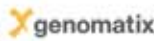
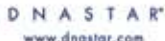
Cambridge Healthtech Institute's Fourth Annual

Next-Generation Sequencing

Data Management

Benefiting from
the Data Bonanza

Corporate Sponsors



Organized by:
Cambridge Healthtech Institute

Healthtech.com/Sequencing

Letter from the editor



Histone modifications, differential methylation, microRNAs — all three work to regulate the genome's content in their own ways. Whether repressing the expression of certain genes or physically blocking interactions between them, miRNAs and epigenetic marks are forces to be reckoned with. For those researchers who are heeding the call of the genomic regulators, it can be work enough just to stay on top of the current technologies, let alone apply them to answer research questions related to epigenetic modifications and miRNA-mediated expression.

For that, *Genome Technology* has again

called on the experts to resolve your technical, planning phase quandaries. For which applications are arrays better suited to an experiment than sequencing? When using high-throughput sequencing for genome-wide methylation-mapping, what's the best targeted capture approach and optimal depth-of-coverage? Is that the same for *Arabidopsis* and *Drosophila*?

In the pages that follow, academic researchers and core lab directors share tips for getting at genomic regulation with ease and precision. Still need more information on miRNAs and epigenetics? Be sure to consult the additional resources at the end of this guide for the most recent methods papers and Web tools in the field.

— Tracy Vence

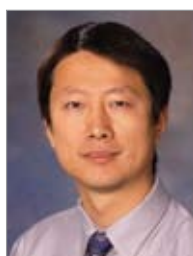
Index of experts

Many thanks to our experts for taking the time to contribute to this technical guide, which would not be possible without them.



Wei Wang

CORNELL UNIVERSITY
LIFE SCIENCES CORE
LABORATORIES CENTER



Wei Li

BAYLOR COLLEGE
OF MEDICINE



Peng Jin

EMORY UNIVERSITY
SCHOOL OF MEDICINE



Expand Your Expertise. Validate Your Experience.

AACC Certificates in Clinical Laboratory Science

Laboratory Management

- ▶ Basic Principles and Architecture of Laboratory Information Systems
- ▶ Clinical Laboratory Leadership and Management *Best Seller!*
- ▶ Financial Management
- ▶ Preparing the Laboratory for Testing
- ▶ Regulatory Affairs
- ▶ Statistical Methods for Clinical Laboratorians

Laboratory Testing

- ▶ Fundamentals of Molecular Pathology
- ▶ Improving Outcomes Through Point-of-Care Testing
- ▶ Laboratory Support for Diabetes Testing
- ▶ Overview of Point of Care
- ▶ Point-of-Care Specialist *Best Seller!*

Laboratory Technology

- ▶ Using Tandem Mass Spectrometry in the Clinical Laboratory *Best Seller!*

Discounts are available for members and groups.

Online education from **AACC**

Learn more at
**aacc.org/
events**

Q1

Which histone modification-mapping techniques do you use, and why?

We utilize histone ChIP-seq approach, which could provide us the best coverage so far.

— Peng Jin

We normally use ChIP-seq because it is the most popular and cost-effective technol-

ogy to get the genome-wide histone modification data.

— Wei Li

Q2

Which genome-scale methylation mapping techniques do you use, and why?

We have been using the microarray-based HELP — HpaII tiny fragment enrichment by ligation-mediated PCR — assay for genome-wide DNA methylation screening. Basically, DNA methylation-dependent restriction digestion patterns are characterized on high-density microarrays to infer the methylation state of the restriction sites. Obviously, this assay interrogates only a fraction of all the potential DNA methylation sites, and it also tends to be susceptible to technical variations in the sample processing, but it can quickly screen large numbers of samples at low cost

and survey sites genome-wide. For validation of the methylation results from the HELP assay, we apply the Sequenom MassArray EpiTYPER assay on small numbers of selected sites. The Sequenom assay is also high-throughput and cost-effective on a large number of samples.

— Wei Wang

We are developing our own approaches right now, given the problems with commonly used approaches. For example, bisulfite sequencing could not distinguish 5mC from 5-hmC, while MeDIP [methyl-

ated DNA immunoprecipitation] could only immunoprecipitate the genomic regions with dense 5mC.

— Peng Jin

We use either whole-genome bisulfite sequencing or reduced representation bisulfite sequencing — RRBS — to profile 5-hmC at single-nucleotide resolution. The unmethylated cytosine is converted to uracil during the bisulfite treatment and sequenced as thymine after PCR amplification, while the methylated cytosine remains unchanged. The methylation ratio is the proportion of remain-

ing cytosines in all the sequencing reads. At current sequencing costs, the former [approach, whole-genome bisulfite sequencing] is still very expensive, while the latter [RRBS] provides an accurate methylation ratio

estimate for the genomic regions of interest in a cost-effective manner. RRBS employs restriction enzyme digestion targeting CCGG, thus focuses on hotspots of epigenetic regulation, such as pro-

motors and CpG islands. By concentrating on a small portion of the genome, RRBS could yield much higher sequencing depth than a whole-genome shotgun approach.

— Wei Li

Q3

Which method do you use to identify and validate microRNA targets?

We are combining both bioinformatic and proteomic approaches. In general, we utilize multiple prediction programs for

initial analyses. We also utilize the SILAC [stable isotope labeling by or with amino acids in cell culture] approach to perform proteomic

analyses [in order] to identify the mRNA targets of any given miRNA.

— Peng Jin

Q4

What measures do you take to ensure reproducibility in your functional analyses of genomic regulation?

To investigate the regulation of gene expression by the epigenetic changes in DNA methylation state, both quantities need to be measured in the same sample to look for the correlation between them among multiple samples. For example, on cancer and normal samples, we can measure the methylation profile

by HELP assay on DNA samples and gene expression pattern by expression microarray on corresponding RNA samples. Therefore, reproducibility of both assays will contribute to the overall reproducibility of this functional analysis of genomic regulation. In my experience, the biggest source of variation is the

sample quality — including both purity and integrity. Stringent quality control needs to be applied on DNA and RNA samples to ensure consistent genomic assay results and good reproducibility. Excluding outlier samples from the genomic data set also improves the overall data quality. Therefore, a larger sample size —

i.e. more biological replicates — is always more desirable. To minimize technical variation in sample processing — due to changes in reagents, personnel, and protocol — the whole study had better be completed in a short period of time, although that is not always feasible for large projects.

— Wei Wang

In general, we utilize multiple biological replicates and technical replicates to ensure the reproducibility of our functional analyses.

— Peng Jin

We check the reproducibility according to NIH ENCODE and Roadmap Epigenome standards

(i.e. for DNA methylation and RNA-seq, the Pearson correlation coefficient needs to be greater than 0.9 between replicates; for ChIP-seq, 80 percent of the top 40 percent of targets from one replicate need to lay within the list from the other replicate and vice versa).

— Wei Li

Q5

When using high-throughput sequencing, how do you balance coverage versus cost for any given experiment?

Coverage needs to meet the minimum requirement for any given high-performance sequencing experiment to obtain reliable analysis result to achieve the specific goal of study. Otherwise, inconclusive or compromised results due to insufficient coverage can actually waste the expense and effort of the study. Under a fixed budget, the sweet spot between the number of samples and the per-sample coverage needs to be determined to answer the particular biological questions. A pilot study, in silico simulation, and the costless literature review will be helpful in the stage of study design. My

inclination is to run fewer samples to guarantee sufficient coverage, and then expand the study to include more samples when more funding becomes available. There are different ways to reduce the cost, [such as] for example, multiplexing samples in the same sequencing run to accurately reach the desirable coverage level, [as does] running biological replicate samples instead of technical replicates.

— Wei Wang

It will depend on the type of assays we are performing. For a resequencing project, we would need to get enough coverage of

the interest regions. For a ChIP-seq project, we typically need to get enough reads to map the peaks with statistical significance. With the [falling] cost of high-throughput sequencing, sufficient coverage would be more important for consideration.

— Peng Jin

“Insufficient coverage can actually waste the expense and effort of the study.”

— Wei Wang

Q6

What are your protocols for data storage and sharing?

Our next-gen sequencers are connected to a small local cluster with many hard drives for temporary network storage of the raw sequencing data files produced. After primary and secondary data analysis, deliverable sequence read data files are transferred to a file server connected to a local high performance computer cluster. The raw sequencing data files are archived to tapes and removed from the small cluster for temporary storage after certain period of time. Customers are promoted to download the read files from the file server via secure FTP links sent in the e-mail notification after the sequencing run. By this means, customers have the freedom to share their read data with any collaborator by forwarding the secure links, and they can have quick access to the read data when they use

the local high-performance cluster for data analysis. This is also a cost-effective solution for sequencing read file distribution.

— Wei Wang

We are currently utilizing the server at our department for storage, mainly due to the availability. We would like to utilize cloud [computing] for future data storage and sharing.

— Peng Jin

We have, in total, [more than] 50 terabytes of high-speed disk storage, which is located in a dedicated server room and maintained by a senior systems administrator. Data are backed up daily and mirrored to a similar disk storage system in an off-site, secured data center. After rigorous verifications, all the raw and processed data [are converted to their] standard formats, with the proper metadata, and will be deposited in

the NCBI Gene Expression Omnibus and Short Read Archive, following established procedures in my lab. We plan to release the data after publication, or one year after data generation regardless of the publication status. In order to provide a uniform platform to facilitate the sharing and comparison of our data, we will establish annual data freezes [in which we create] a snapshot of all data sets that have been made available by the freeze date.

— Wei Li

“We would like to utilize cloud [computing] for future data storage and sharing.”

— Peng Jin

Genomic Regulation Grants

GRANT OPPORTUNITIES

Organization: National Institutes of Health, National Institute on Drug Abuse

Award: Size and duration will vary according to the nature and scope of the proposed research.

Details: This grant will support research aimed at functional genetics, epigenetics, and non-coding RNAs in drug addiction. The NIH encourages basic genomics research into the fundamental biological mechanisms underpinning addictive processes, including the functional validation of candidate genes and the elucidation of the molecular pathways and processes they involve.

Contact: Scientific/Research, John Satterlee (satterleej@nida.nih.gov); Financial/Grants Management, Deborah Wertz (dwertz@nida.nih.gov)

Organization: National Institutes of Health, National Cancer Institute

Award: Size and duration will vary according to the nature and scope of the proposed research.

Details: The National Cancer Institute intends to support projects that aim to evaluate methylation profiles, histone modifications, and microRNAs associated with the risk of developing cancer in different populations.

Contact: Scientific/Research, Mukesh Verma (vermam@mail.nih.gov); Financial/Grants Management, Crystal Wolfrey (wolfrey@mail.nih.gov)

Organization: National Science Foundation

Award: Size and duration will vary according to the nature and scope of the proposed research. awards funded in FY 2010 ranged from \$634,846 for five years to \$9,946,315 for four years.

Details: NSF intends to support plant genomics projects that aim to address major unanswered questions in plant biology on a genome-wide scale, and is accepting proposals at all scales — from single-investigator projects through multi-institution projects.

Contact: Diane Jofuku Okamuro (dbjofuku@nsf.gov)

e^xl
pharma PROUDLY PRESENTS



Pharmaceutical Research Collaborations Summit

Establishing Value and Ensuring Successful Operations when Choosing to Collaborate with Strategic Partners, Public-Private Partnerships, Universities, and Open Innovation Programs

JULY 26 - 27, 2011
RADISSON HOTEL BOSTON
BOSTON, MA

SAVE 15% OFF
the Standard
Registration Rates!
Simply Enter
GENOME when
registering!

FEATURING PRESENTATIONS FROM LEADING UNIVERSITIES AND PHARMACEUTICAL COMPANIES:

CREATING AN EXTERNAL PORTFOLIO CASE STUDY

Sourcing, Developing, and Funding New Molecules Through Outside Sources

Aaron Schacht, Executive Director, Global R&D, **ELI LILLY**

PARTNERING WITH ACADEMIC INSTITUTIONS

How Pharmaceutical Companies can Benefit from and Contribute to the Expansive Research Being Done at Academic Institutions in the Form of Private-Public Partnerships

Reid Leonard, Executive Director, External Licensing and Scientific Affairs, **MERCK**

PRE-COMPETITIVE CASE STUDY

The Open Pharmacological Space Project

Bryn Williams-Jones, eBiology Group Leader, **PFIZER**

DON'T MISS THE EXECUTIVE PANEL:

COLLABORATIONS OVERVIEW PANEL: Leaders from Big Pharma, Academia, Smaller Biotechs, Non-Profits, and Government Discuss Collaborations Between Groups

- How each stakeholder can benefit in a collaborative pharmaceutical research partnership
- Evaluating the public health concerns that drive private-public partnerships
- Economic factors affecting how industry and academia conduct biomedical research

Alan Lamont, Director, SPBD, Science and Technology Licensing, **ASTRAZENECA**

Brent Bankosky, Senior Director, Global Licensing & Business Development, **TAKEDA**

Lita Nelsen, Director, Technology Licensing Office, **MASSACHUSETTS INSTITUTE OF TECHNOLOGY**

Tim Dolan, Director, Business Development, **ELI LILLY**

To Register Call 866-207-6528 or Visit Us
at www.exlpharma.com/collaborationssummit

List of resources

For as many recognized mechanisms of genomic regulation that exist, there are at least twice as many approaches one can take to study each. More still are the options for bioinformatics analysis from which to choose. Here's a selection of recent methods papers, standby Web tools, and must-attend meetings in the field.

PUBLICATIONS

- Alexiou P, Manolis M, Hatzigeorgiou AG. (2011). **Online resources for microRNA analysis.** *Journal of Nucleic Acids Investigation*. Epub: doi 10.4081/jnai.2011.e4.
- Bussotti G, Raineri E, Erb I, Zytnicki M, Wilm A, Beaudoin E, Bucher P, Notredame C. (2011). **BlastR — fast and accurate database searches for non-coding RNAs.** *Nucleic Acids Research*. Epub: doi 10.1093/nar/gkr335.
- Chen L, Wu G, Ji H. (2011). **hmChIP: a database and Web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data.** *Bioinformatics*. 27(10): 1447-1448.
- Chen Y, Meyer CA, Liu T, Li Wei, Liu JS, Liu XS. (2011). **MM-ChIP enables integrative analysis of cross-platform and between-laboratory ChIP-chip or ChIP-seq data.** *Genome Biology*. 12: R11.
- Deorowicz S, Grabowski S. (2011). **Compression of DNA sequence reads in FASTQ format.** *Bioinformatics*. 27(6): 860-862.
- Elefant N, Berger A, Shein H, Hofree M, Margalit H, Altuvia Y. (2011). **RepTar: a database of predicted cellular targets of host and viral miRNAs.** *Nucleic Acids Research*. 39 (Suppl1): 188-194.
- Fejes AP, Khodabakhshi AH, Birol I, Jones SJ. (2011). **Human variation database: an open-source database template for genomic discovery.** *Bioinformatics*. 27(8): 1155-1156.
- Francesconi M, Jelier R, Lehner B. (2011). **Integrated genome-scale prediction of detrimental mutations in transcription networks.** *PLoS Genetics*. 7(5): e1002077.
- Fritz MH, Leinonen R, Cochrane G, Birney E. (2011). **Efficient storage of high-throughput DNA sequencing data using reference-based compression.** *Genome Research*. 21: 734-740.
- Gu J, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. (2011). **Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling.** *Nature Protocols*. 6: 468-481.
- Hummel M, Bonnin S, Lowy E, Roma G. (2011). **TEQC: an R package for quality control in target capture experiments.** *Bioinformatics*. 27(9): 1316-1317.
- Krueger F, Andrews SR. (2011). **Bismark: a flexible aligner and methylation caller for bisulfite-seq applications.** *Bioinformatics*. 27(11): 1571-1572.
- Lutsik P, Feuerbach L, Arand J, Lengauer T, Walter J, Bock C. (2011). **BiQ Analyzer HT: locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing.** *Nucleic Acids Research*. Epub: doi 10.1093/nar/gkr312.

Muiño JM, Hoogstraat M, van Ham RC, van Dijk AD. (2011). **PRI-CAT: a web-tool for the analysis, storage and visualization of plant ChIP-seq experiments.** *Nucleic Acids Research*. Epub: doi 10.1093/nar/gkr373.

Pardo CE, Carr IM, Hoffman CJ, Darst RP, Markham AF, Bonthrom DT, Kladde MP. (2011). **MethylViewer: computational analysis and editing for bisulfite sequencing and methyltransferase accessibility protocol for individual templates (MAPit) projects.** *Nucleic Acids Research*. 39 (1): e5.

Qin J, Li MJ, Wang P, Zhang MQ, Wang J. (2011). **ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/in-direct targets of a transcription factor.** *Nucleic Acids Research*. Epub: doi 10.1093/nar/gkr332.

Shankaranarayanan P, Mendoza-Parra MA, Walia M, Wang L, Li N, Trindade LM, Grone-meyer H. (2011). **Single-tube linear DNA amplification (LinDA) for robust ChIP-seq.** *Nature Methods*. Epub: doi 10.1038/nmeth.1626.

Shen Y, Song R, Pe'er I. (2011). **Coverage tradeoffs and power estimation in the design of whole-genome sequencing experiments for detecting association.** *Bioinformatics*. Epub: doi 10.1093/bioinformatics/btr305.

Vavouri T, Lehner B. (2011). **Chromatin organization in sperm may be the major functional consequence of base composition variation in the human genome.** *PLoS*

Genetics. 7(4): e1002036.

Wang C, Zhang D. (2011). **A novel compression tool for efficient storage of genome resequencing data.** *Nucleic Acids Research*. 39(7): e45.

Zisoulis DG, Yeo GW, Pasquinelli AE. (2011). **Comprehensive identification of miRNA target sites in live animals.** *Methods in Molecular Biology*. 732: 169-185.

WEB SITES

ChromDB

<http://www.chromdb.org/>

MethDB

<http://www.methdb.de/>

miRBase

<http://www.mirbase.org/>

miRDB

<http://mirdb.org/miRDB/>

miRNA Target Database

http://www.ncrna.org/KnowledgeBase/link-database/mirna_target_database

NCBI Gene Expression Omnibus

<http://www.ncbi.nlm.nih.gov/geo/>

NCBI Short Read Archive

<http://www.ncbi.nlm.nih.gov/Traces/sra>

NHGRI Histone Sequence Database

<http://research.nhgri.nih.gov/histones/>

PubMeth

<http://www.pubmeth.org/>

TargetScan

<http://genes.mit.edu/targetscan/index.html>

CONFERENCES

Epigenetics: Mechanisms, Development, and Disease

Gordon Research Conferences

Aug 7-12, 2011

Easton, Mass.

Epigenetics Europe

Select Biosciences

Sep 8-9, 2011

Munich, Germany

RNAi & miRNA Europe

Select Biosciences

Sep 8-9, 2011

Munich, Germany

Epigenomics of Common Diseases

Wellcome Trust

Sep 13-16, 2011

Hinxton, UK

EMBO Workshop: Histone Variants & Genome Regulation

European Molecular Biology Organization

Oct 12-14, 2011

Strasbourg, France

INSERM Workshop: High-Throughput Approaches in Epigenomics

Institut National de la Santé et de la Recherche Médicale

Oct 10-12, 2011

Bordeaux, France

INSERM Workshop: Bioinformatics Approaches to Decipher Genome Regulation

Institut National de la Santé et de la Recherche Médicale

Oct 12-14, 2011

Bordeaux, France

MicroRNAs Europe 2011

GeneExpression Systems

Nov 1-2, 2011

Cambridge, UK

Genome Informatics

Cold Spring Harbor Laboratory, Wellcome Trust

Nov 2-5, 2011

Cold Spring Harbor, NY

X CRG Annual Symposium: Computational Biology of Molecular Sequences

Centre for Genomic Regulation

Nov 10-11, 2011

Barcelona

Next-Generation Sequencing

Congress Europe

Oxford Global Conferences

Nov 14-15, 2011

London

EuroEpiStem: European Epigenomics & Stem Cells

GeneExpression Systems

Nov 21-22, 2011

Paris

RNAi Asia

Select Biosciences

Nov 22, 2011

Singapore

Chromatin: Structure & Function

Abcam

Dec 5-8, 2011

Aruba

Epigenomics

Keystone Symposia

Jan 12-22, 2012

Keystone, Colo.

Gene Silencing by Small RNAs

Keystone Symposia

Feb 7, 2012

Keystone, Colo.

We scan Nature

(and a few others)

so you don't have to.

**If you only have five minutes
a day, here's what you
need to read.**

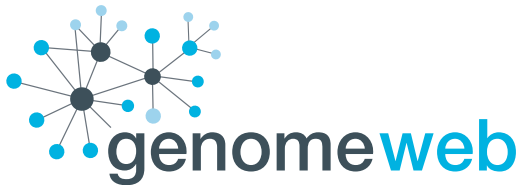
**We comb through more than 60 science blogs
to recommend the most interesting posts.**

**Weekly reports of the news and papers
in Nature and Science, plus highlights from journals
such as PLoS Biology and PNAS.**

**Can't get to every major newspaper and magazine?
We scan the mainstream media to keep you
up on the news you need.**

**Register for your free Daily Scan Bulletin
at www.genomeweb.com**





The research community has a new place to gather.
Check out our fresh look and features at genomeweb.com.

Can't wait a month for *Genome Technology*?
Get your daily fix at the new GenomeWeb.



Check out some of our new features:

+ Expanded Access

Most academic and government researchers can access all of GenomeWeb's premium content at no cost. Login with your workplace e-mail to qualify.

+ More Content

Everyone has complete free access to GenomeWeb Daily News, the Daily Scan, Genome Technology, and more.

+ Easy Login

Login just once and navigate easily to all of your GenomeWeb content.

+ Magazine E-mail

Get the *Genome Technology* table of contents by e-mail every month. It also includes the PDF edition of the magazine and tech guides.