FDA U.S. FOOD & DRUG ADMINISTRATION
NATIONAL CENTER FOR TOXICOLOGICAL RESEARCH

# Division of Bioinformatics and Biostatistics

Presented by:

Weida Tong, Ph.D.
National Center For Toxicological Research
U.S. Food and Drug Administration

***Disclaimer:*** *The information in these materials is not a formal dissemination of information by FDA and does not represent agency position or policy.*

# Division Staff

- Four branches (44 Full Time Employees including vacancies):
  - Bioinformatics Branch: ***Research*** centric (9 FTEs)
  - Biostatistics Branch: ***Research*** + ***Support*** + ***Service*** (9 FTEs)
  - R2R Branch: focusing on ***Research*** and ***Support*** (9 FTEs) – Newly established (2017)
  - Scientific Computing Branch (SCB): ***Service*** centric (17 FTEs)

- Immediate office: 3 administrators + one senior advisor

- 5 Postdoctoral fellows + 2 Students

- The Division is multidisciplinary:
  - IT specific expertise: software engineers, programmers, system administrators, etc
  - Research scientists: Bio/chemo-informatics, (bio)statistics, computational chemistry, and biology

# Three Functions

- **Service**: legacy activities (*will not be presented*)
  - Scientific Computing Branch (17 FTEs): Computer Center (135 servers, PB of storage, HPC cluster, working with 12 FTEs from OIM)
  - Biostatistics Branch (4 FTEs) to support NTP: statistical support
  - Part of center-wide infrastructure and investment
- **Support**: bioinformatics and biostatistics support (*will be presented briefly*)
  - Mainly in the R2R and Biostatistics Branches
  - Within NCTR – Engage two types of support: "**Committed**" tasks and "**Requested**" activity
  - Beyond NCTR – Develop and maintain tools for regulatory review process (e.g., CDER)
- **Research**: PI and peer-reviewed (*will be the focus*)
  - Mainly in the Bioinformatics (9 FTEs) and Biostatistics (4 FTEs) Branches

# Support – Within NCTR

- **"Committed" tasks**: Establish data analysis environment, manage commercial and in-house software tools, and conduct training courses
  - Next-generation sequencing (NGS) data: (1) Implement Galaxy Platform, (2) Manage CLC Genomics Workbench, (3) Maintain HPC, and (4) Coordinate/lead the NGS club
  - ArrayTrack for microarray data management, analysis and interpretation
  - Provide the training to use these tools

- **"Requested" activity**: Provide data analysis capability and method development requested by PIs from other divisions
  - Sequencing data analysis (NeuroTox, DGMT, DBT) for microRNAs, methylation data (epigenetics), RNA-seq and DNA-seq
  - Integrated analysis of multiple sources of omics data (NeuroTox)
  - Image data analysis (e.g., nano-core)
  - Biostatistical support (DBT and DSB)
  - Data management software development (Microbiology)
  - FDALabel: FDA drug labeling database (OSC, CDER/OND, and CDER/OCS)

# Support and Collaborate with Other FDA Centers

- **CDER:** improve regulatory submission, scientific review and knowledge management
  - IND template to standardize the IND data submission and management (CDER/OCS)
  - Support DASH (Data Analysis Search Host) Tool
    - DASH tracks approval of NDAs/BLAs and the progression from INDs to NDAs or BLAs (CDER/OTS knowledge management team)
    - Objectives: improve DASH performance (user-interface and DB) and increase its utility by integrating with other review data
  - Text mining: Approval letters, patients narratives, OND regulatory documents (Meeting Minutes), PharmTox documents, and DASH text
- **CBER:** Use NGS to assess safety of engineered therapeutic biologics
- **CTP:** Use molecular modeling to assess addiction potential of tobacco constituents

# Enhance “Support” Function

- The R2R framework (***R**esearch-**To**-**R**eview and Return*)
  - A framework for collaboration & forming/reinforcing linkages with product Centers
  - Co-managed by NCTR (DBB) and CDER (OCS)
  - Four stages of agile process governing R2R:
    - Stage 1 (**Exploration**): define and prioritize potential projects and understand the requirements via periodic F2F meetings and video conferences
    - Stage 2 (**Assessment**): prototyping and testing to determine “go” or “no-go”
    - Stage 3 (**Execution**): develop phased program plan with deliverables, timeline, and milestones
    - Stage 4 (**Completion**): determine follow-on action: (1) proceeded for production, (2) new projects derived from expansion/extension, or (3) work finished
- Projects governed by R2R – FDALabel, DASH, IND template, Galaxy, Visualization tools (e.g., HCA and PCA), etc
- Established a new branch, R2R branch (9 FTEs), in 2017

# R2R Branch (Established in 2017)

- Focused on “support” to both NCTR and FDA centers:
  - Formalized R2R oversight
  - Support new data streams (omics, HTP/HCS, imaging):
    - HPC, software platforms (e.g., Galaxy), center-wide software licenses
    - Algorithm/pipeline implementation/development, difficult data analysis (e.g., bioimaging and NGS data, and integrated analysis)

- Benefits: Easier priority-setting and coordination of the “support” function with an efficient way of utilizing resources and training personnel
  - The support function is coordinated and supervised.
  - Job responsibility and career development path for the R2R staffs are clarified.
  - The oversight of current projects and identifying future needs will be improved.

# Division Mission

- **Service/Support**
  - To provide research and regulatory support to NCTR and FDA scientists in bioinformatics, biostatistics, and scientific computing.
  - To ensure that the division's activities relate to FDA's review process, our linkages with product centers continue to be strengthened, and our capabilities evolve to meet the current and future needs of FDA.

- **Research**
  - To conduct integrative bioinformatics and biostatistics research to support FDA's mission of improving the safety and efficacy of FDA-regulated products.

# Global Leadership and Outreach

- **Global leadership and outreach**
  - FDA-specific activities: FDA working group participation (e.g., co-chair for LiverTox WG, GWG, GnG, SCB, modeling&simulation …)
  - Community-specific activities:
    - SOT (and local chapter), AAPS, MAQC Society
    - An active role in Global Coalition for Regulatory Science Research (GCRSR)
  - SAB committee: projects in both US and Europe

- **Regional activities**
  - Arkansas Bioinformatics Consortium (AR-BIC):
    - Established in 2014; consists of the Arkansas major universities plus NCTR; The division formulated the concept and facilitated its establishment and development
    - Participated in organizing three annual AR-BIC conferences
  - MidSouth Computational Biology and Bioinformatics Society (MCBIOS): A Past-President with two Board of Directors

# DBB Research (13 FTEs)

*Top Three Accomplishments During the Last 5 Years*

# Accomplishment #1: MAQC

- MicroArray Quality Control (**MAQC**): An FDA-led consortium effort to assess technical performance and application of emerging technologies for safety evaluation and clinical application

- Completed MAQC-I (microarrays) by 2006 and MAQC-II (microarrays and GWAS) by 2010
  - ~20 publications, 8 in *Nat Biotechnol*, 2 are among top 10 most-cited paper in *Nat Biotech* in the past 20 years

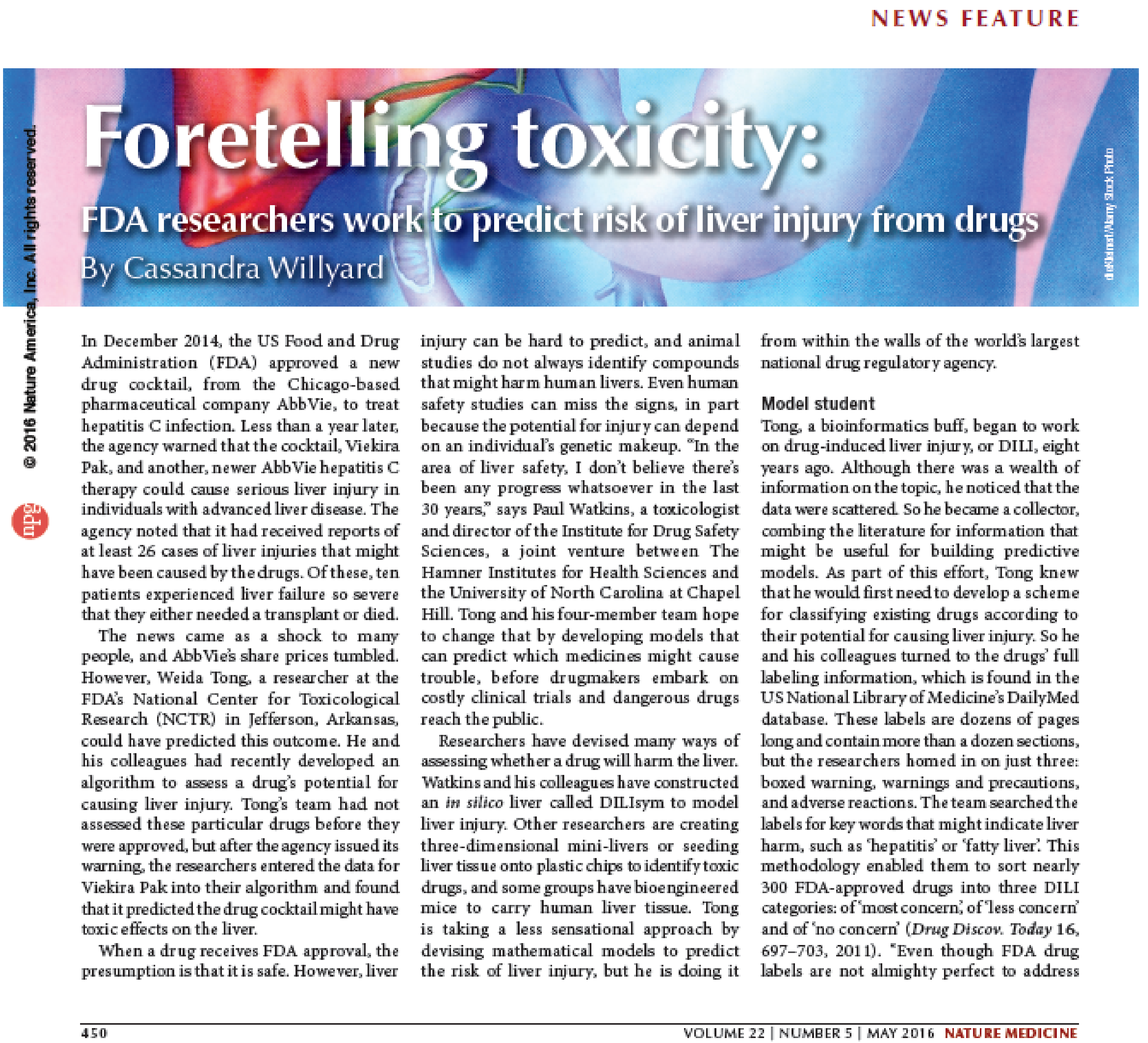
- Completed MAQC-III (RNA-sequencing) by 2014; also known as Sequencing Quality Control (SEQC)
  - More than 180 participants from 73 organizations
  - Generated > 10Tb data; represented ~6% data in GEO (Jun, 2014)
  - 10 Manuscripts: 3 in *Nat Biotechnol*, 2 in *Nat Commun*, 3 in *Sci Data*, 2 in *Genome Biology*

# Accomplishment #2: LTKB

- Liver Toxicity Knowledge Base (LTKB): A knowledgebase system for predicting drug-induced liver injury (DILI) in humans, which involves
  - Curating a broad range of data associated with marketed drugs
  - Developing predictive models that can be used individually or in combination to predict DILI potential

- Accomplishments:
  - Constructed a database containing diverse datasets (e.g., drug-centric data, genomics, *in vitro* assays, etc.) freely available for public use
  - Developed six predictive models for DILI in humans (one model is being evaluated for the FDA review process)

- Six publications in the last two years (2016 and 2017):
  - *Gastroenterology* (IF=18), *Hepatology* (IF=11), *Journal of Chemical Information and Modeling* (IF=3.5), *Drug Discovery Today* (IF=6)



# Foretelling toxicity:

## FDA researchers work to predict risk of liver injury from drugs

By Cassandra Willyard

dieKleinert/Alamy Stock Photo

In December 2014, the US Food and Drug Administration (FDA) approved a new drug cocktail, from the Chicago-based pharmaceutical company AbbVie, to treat hepatitis C infection. Less than a year later, the agency warned that the cocktail, Viekira Pak, and another, newer AbbVie hepatitis C therapy could cause serious liver injury in individuals with advanced liver disease. The agency noted that it had received reports of at least 26 cases of liver injuries that might have been caused by the drugs. Of these, ten patients experienced liver failure so severe that they either needed a transplant or died.

The news came as a shock to many people, and AbbVie's share prices tumbled. However, Weida Tong, a researcher at the FDA's National Center for Toxicological Research (NCTR) in Jefferson, Arkansas, could have predicted this outcome. He and his colleagues had recently developed an algorithm to assess a drug's potential for causing liver injury. Tong's team had not assessed these particular drugs before they were approved, but after the agency issued its warning, the researchers entered the data for Viekira Pak into their algorithm and found that it predicted the drug cocktail might have toxic effects on the liver.

When a drug receives FDA approval, the presumption is that it is safe. However, liver injury can be hard to predict, and animal studies do not always identify compounds that might harm human livers. Even human safety studies can miss the signs, in part because the potential for injury can depend on an individual's genetic makeup. "In the area of liver safety, I don't believe there's been any progress whatsoever in the last 30 years," says Paul Watkins, a toxicologist and director of the Institute for Drug Safety Sciences, a joint venture between The Hamner Institutes for Health Sciences and the University of North Carolina at Chapel Hill. Tong and his four-member team hope to change that by developing models that can predict which medicines might cause trouble, before drugmakers embark on costly clinical trials and dangerous drugs reach the public.

Researchers have devised many ways of assessing whether a drug will harm the liver. Watkins and his colleagues have constructed an *in silico* liver called DILIsym to model liver injury. Other researchers are creating three-dimensional mini-livers or seeding liver tissue onto plastic chips to identify toxic drugs, and some groups have bioengineered mice to carry human liver tissue. Tong is taking a less sensational approach by devising mathematical models to predict the risk of liver injury, but he is doing it from within the walls of the world's largest national drug regulatory agency.

### Model student

Tong, a bioinformatics buff, began to work on drug-induced liver injury, or DILI, eight years ago. Although there was a wealth of information on the topic, he noticed that the data were scattered. So he became a collector, combing the literature for information that might be useful for building predictive models. As part of this effort, Tong knew that he would first need to develop a scheme for classifying existing drugs according to their potential for causing liver injury. So he and his colleagues turned to the drugs' full labeling information, which is found in the US National Library of Medicine's DailyMed database. These labels are dozens of pages long and contain more than a dozen sections, but the researchers homed in on just three: boxed warning, warnings and precautions, and adverse reactions. The team searched the labels for key words that might indicate liver harm, such as 'hepatitis' or 'fatty liver'. This methodology enabled them to sort nearly 300 FDA-approved drugs into three DILI categories: of 'most concern', of 'less concern' and of 'no concern' (*Drug Discov. Today* 16, 697–703, 2011). "Even though FDA drug labels are not almighty perfect to address

# Accomplishment #3: Rare Disease

- **Mission:** Study of rare disease with bioinformatics
  - ~7000 reported rare diseases with only ~500 therapeutic options
  - One third of FDA approved drugs over the past 10 years for rare diseases

- **Goals:** *In silico* drug repositioning to exploit marketed drugs for safer and affordable alternatives for rare diseases
  - Deciphering miRNA transcription factor feed-forward loops to identify drug repurposing candidates for cystic fibrosis
  - Potential reuse of oncologic drugs for the treatment of rare diseases

- **Publications**
  - 2 publications in Trends in Pharmacological Sciences (IF=10.15), Drug Discovery Today (IF=6), Genome Medicine (IF=7)

# Examples of Current Projects

Will be explained in more detail

- Sequencing Quality Control Phase 2 (SEQC2)
- Liver Toxicity Knowledge Base (LTKB)
  - Evaluation and developement of DILI predictive models for INDs
  - Study of drug-host interaction with electronic health records (collaborating with VA)
  - Hepatotoxicity database for herbal medicine and dietary supplement (funded by OWH and in collaboration with CDER and CFSAN)
- Endocrine Disruptor Knowledge Base (EDKB) (~50 publications for the last 20yrs)
- Big data methodologies and applications
  - Data integration, e.g., study of rare diseases with bioinformatics
  - Text mining, e.g., topic models for analysis FDA documents
  - Image analysis, e.g., develop an image-based recognition system for food contamination with bugs using machine learning and deep learning (collaborated with ORA)

# SEQC2: Overview

- Focused on DNA-seq
  - The 4th MAQC/SEQC consortium project
  - A natural extension of SEQC1; Focused on whole genome sequencing (WGS) and target gene sequencing (TGS)
  - Assessing QC, reproducibility, and bioinformatics

- Milestones
  - 2014. 11 ~ 2015.03 – Discussed among the SEQC leadership team; 4 areas of focus were recommended
  - 2015.03 ~ 2015.09 – Established the FDA team (NCTR, CBER, CDRH, CDER, CFSAN, CVM, CTP, OCS)
  - 2015.12 – Communicated with NBT
  - Workshops
    - Kick-off meeting: NIH Campus, Sept 13-14, 2016
    - 2nd workshop: SAS Campus, April 10-12, 2017

# SEQC2: Objectives and Goals

- **Reproducibility**
  - Cross-lab
  - Cross-platform
  - Cross-analysis methods

- **Data Analysis**
  - Benchmark bioinformatics methods (pipeline, variant calls, coverage and joint effect)
  - Towards standard analysis protocol

- **Clinical Relevance (precision medicine)**
  - Target gene sequencing (present)
  - Whole Genome Sequencing (future application)

# SEQC2: Working Groups

## Objective

Develop best practices with recommended standard analysis protocols and quality control metrics for whole genome sequencing (WGS) and targeted gene sequencing (TGS) technologies that will support regulatory science research and precision medicine.

**WG#1: Somatic Mutation**

A comprehensive comparison across various labs, NGS methods, and data analysis protocols using a pair of tumor normal cell lines (ATCC) and FFPE. Over 50 institutes with >160 members participate in data generation and analysis.

**WG#2: Onco-Panel Sequencing**

Assess detection power and across-site and platforms reproducibility for onco-panel sequencing in detection of subclonal mutation and liquid biopsy.

**WG#3: Germline Variants**

Assess the WGS accuracy and reproducibility for variants call by investigating the join effect of reads alignment pipelines, variants call methods and coverage.

**WG#4: Difficult Genes**

Assess the accuracy for some difficult genes that varies significantly due to complexity in their genomic regions (e.g. GC region) with specifically focused on HLA genes.

**WG#7: Epigenetics**

Study of dynamic epigenetic loci and their phased haplotypes (epialleles) with standardized materials, methods, and rigorous benchmarking to enable and improve the epigenetic application to clinical and research projects

# Liver Toxicity Knowledge Base (LTKB)

- **Communicating LTKB models via the FDA LiverTox Working Groups**
  - Rule-Of-Two (RO2) model has been evaluated by the FDA
    - Applied to ~20 cases ranging from IND, NDA to post-market processes
    - Resulted a joint publication with CDER in Gastroenterology (2017)
    - Defined the fit-for-purpose application: RO2 not works well for the drugs with immune responses

- **Further improve LTKB models**
  - Annotating all the drugs approved by the FDA for their likelihood of causing DILI (>1000 drugs)
  - Applying systems approach for an enhanced DILI prediction:
    - Model integration: develop predictive models for each dataset and combine them in a consensus or hierarchical fashion
    - Data integration: integrate diverse data to develop a single model
    - Drug-host integration: Integrate host factors (genetic factors, sex, age, etc) into drug centric predictive models

# Future Directions

- Data science-centric support
  - Increase data analysis support such as imaging and NGS data to improve our ability to deal with big data
  - Advance the R2R program to strengthen our linkage to other centers
- Five-Year plan for DBB research (new imitative):
  - To integrate LTKB models in the review process
  - To leverage the LTKB methodology to study cardiovascular toxicity
  - To continually develop big data analytics (e.g., deep learning, Bayesian approach, data integration)
  - To address issues related to computational reproducibility
  - To conduct crowdsourcing and community-wide project to assess emerging methodologies; e.g., *In Silico* DILI (isDILI) project

# Feedback Requested (1)

- How to efficiently utilize the growing size of the diverse datasets in the public domain to address the FDA issues?
  - Where is the major advancement in systems approach and integrated analysis methodology?
  - What knowledge bases and databases are useful for regulatory uses (besides the ones we have developed for liver toxicity, endocrine disruptors and liver carcinogenicity)?
  - Which datasets can advance *in vitro* and *in silico* biomarkers for human safety (besides the data from ToxCast and Tox21)?
  - How about the electronic health records; will they mature and where you see the challenges?

# Feedback Requested (2)

- What emerging technologies are on the horizon that need our engagement?
  - How to develop the fit-for-purpose metrics to objectively assess these technologies?
  - How should standard reporting mechanisms be developed for traceability and transparency of results from these technologies?

- Given both support and research are critical for projects involving bioinformatics and biostatistics, what might you suggest to encourage research scientists involving more in support/service?

FDA U.S. FOOD & DRUG
ADMINISTRATION