

clinical data

observational & scientific data

computation & visualization

decision support

cceHUB

Sharing, Exploring and Analyzing Data

An Environment for Collaborative Cancer Research

Ann Christine Catlin HUBbub 04-06-2011



Master plan for the Cancer Care Engineering Colorectal Cancer Study

1. Blood Sample Acquisition

Sample Processing, Annotation, Distribution Clinical Patient Data Collection IU Simon Cancer Center

2. **OMIC Laboratory Analysis**

Data & Knowledge Acquisition

Xu Lab Lipidomics IU School of Medicine

Raftery Lab Metabolomics Purdue

Regnier Lab Glycoproteomics Purdue

Bindley Lab Global Proteomics Purdue

Teegarden Lab Vitamin D Purdue

Klaunig Lab Oxidative Stress IU School of Medicine

3. Predictive Modeling

Data Synthesis & Analysis
Knowledge Acquisition
Zhang Group Integrative Models Purdue
Sherer Population-based Models VA Hospital
Chen Biological Network Models IUPUI

4. Visual Analytics

Data Exploration & Analysis
Knowledge Acquisition
Ebert Group PURVAC Purdue

5. <u>Iterative Feedback & Validation</u>

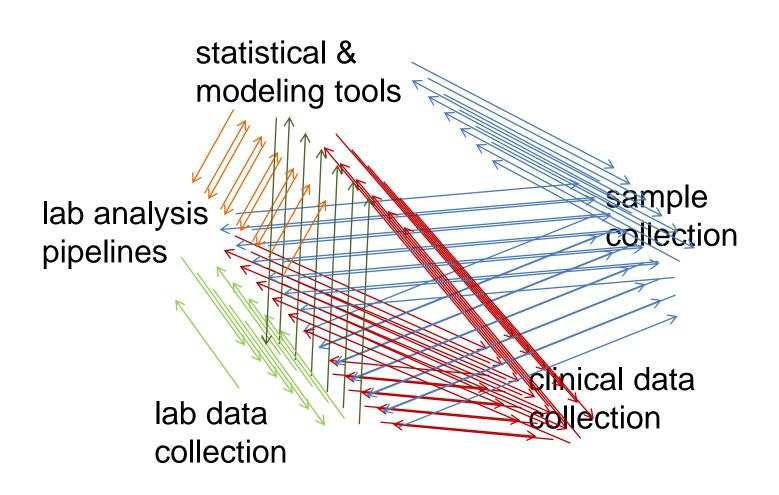
CCE Research Community



molecular signatures for colorectal cancer that predict susceptibility, treatment response and ultimate treatment outcome

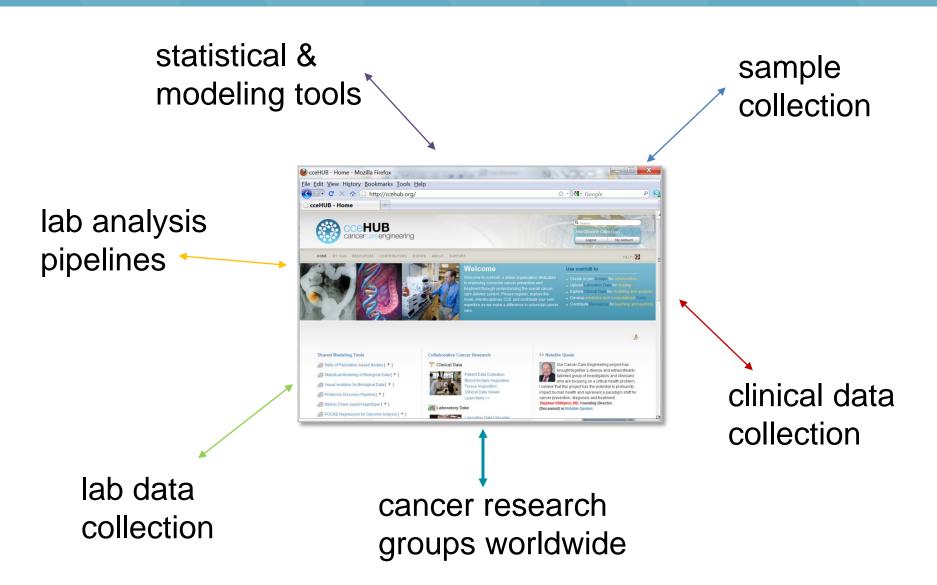


Sharing data, tools, analysis & knowledge



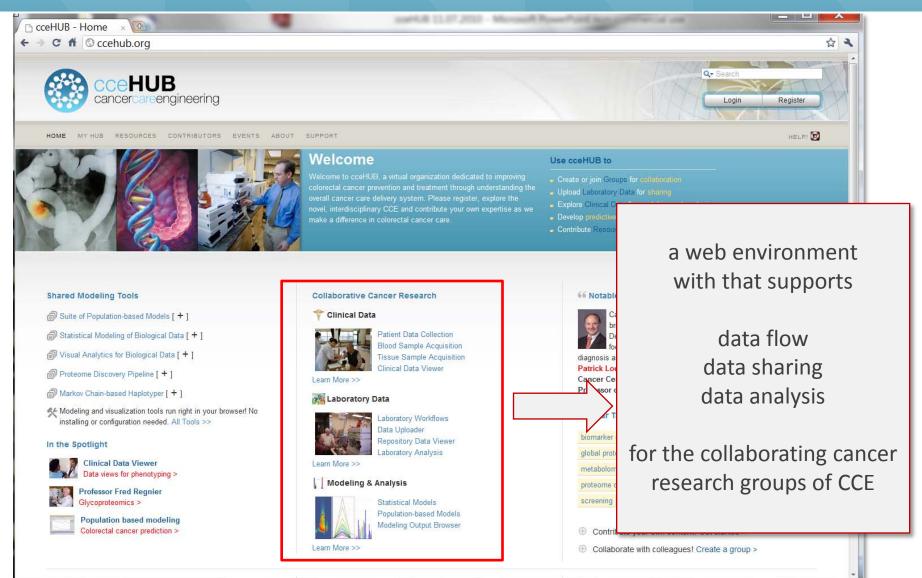


A single portal: sharing data, tools, analysis & knowledge





A single portal: sharing data, tools, analysis & knowledge





Support for clinical data



Clinical Research Team and Physicians

Data

Workflow

Data contribution from clinical team

Patients Diagnosis, Treatments, Surgeries, Lifestyle, Diet, Demographics, ...

Samples Collection, Processing, Protocols, Distribution, Tracking, ...

Automatic Metadata Processing

Sample Data
Patient Data
Clinical Metadata

cceHUB Database



Support for clinical data



7



Patient Data Collection
Blood Sample Acquisition
Tissue Acquisition
Clinical Data Viewer
Learn More >>

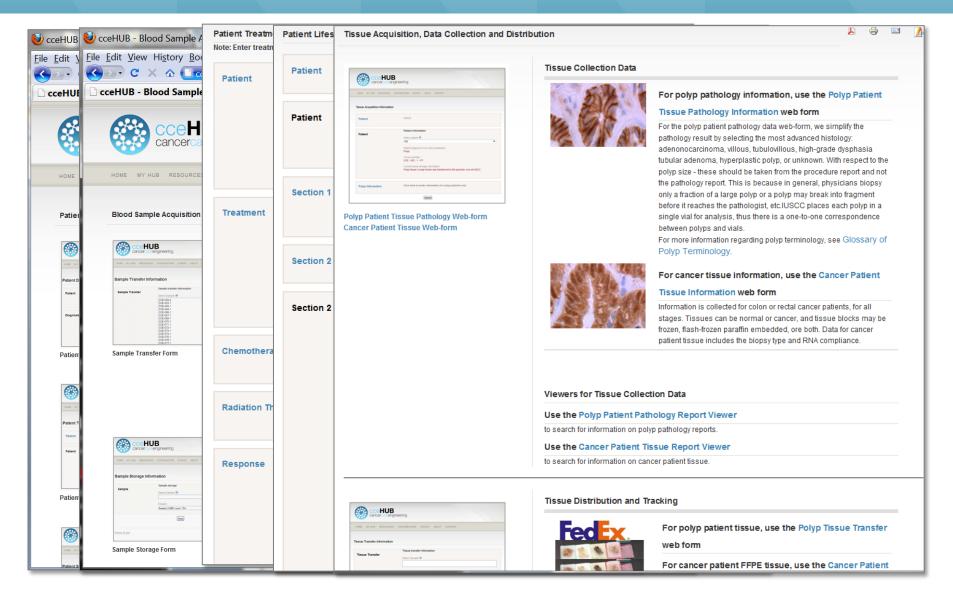
Clinical Data Flow

- nightly pull from hospital e-records
- patient data collection
- sample tracking
- data annotation
- clinical data archive
- blood sample bio-repository
- patient and sample linkage
- data viewing
- data search, filter & explore

a lata

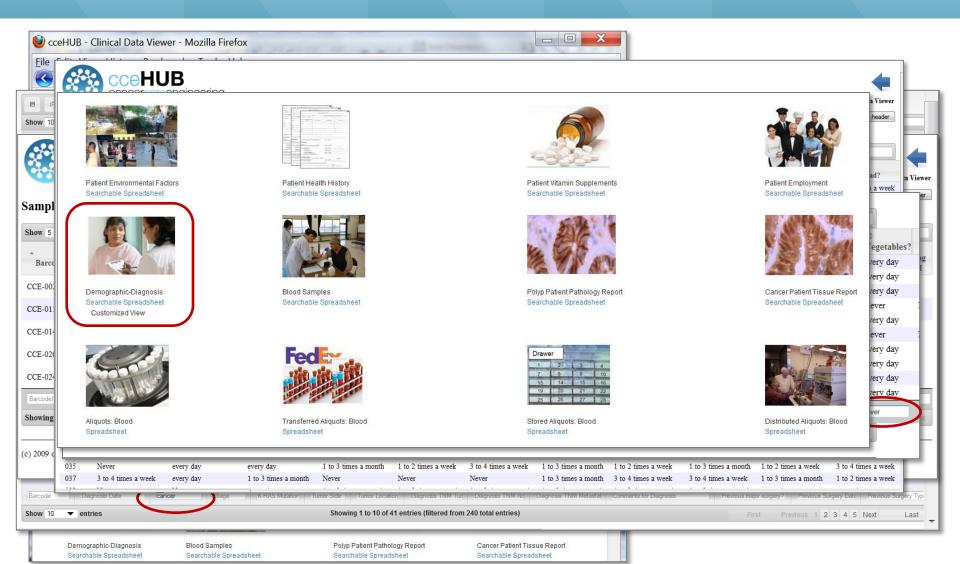


Clinical and sample data collection & processing





Clinical data views





Clinical data: some stats

Database	Total	Diagnosis	Lifestyle	Cancer/Polyp	Treatment
	Patients	% Data	% Data	Patients	% Data
patients	240	100%	70%	41 / 92	100%

First patient CCE001 enrolled on 04/02/2009 (the day cceHUB went live)

Most recent patient CC285 enrolled on 02/15/2011

Most recent data: neoadjuvant chemoradiation treatment for patient CCE156 on 04/02/2011 Maximum patients enrolled on a single day 09/23/2009 = 9

web-forms to track patient and sample data flow : 12

accesses to clinical data viewer 04/02/2009 - 05/25/2010 : > 15,000

Database	Total Samples	Total Aliquots	Sample Tracking Web-forms	# instances cceHUB used to find missing aliquot
samples	267	5073	sample processing sample transfer sample storage sample distribution	52 (we track sample barcodes, location, entry person, entry date)



Support for laboratory data



Tricosanoic

Research Labs Metabolomics, Lipidomics, Global Proteomics, Glycoproteomics, Vitamin D, Oxidative Stress, Genomics

Lab Workflow Knowledge Sample-Dataset tracking Massive instrument- generated datasets Data Upload **Clinical Data** Lab Knowledge Base **Repository Metadata**

cceHUB Lab Instrument Data Repository

cceHUB **Database**



Support for laboratory data







Laboratory Data Uploader
User Collections Uploader
Repository Data Viewer
Track Laboratory Analysis 2011, 2010
Learn More >>

Laboratory Data Flow

- "knowledge base" resources
 (protocols, sample preparation, instruments, standards, file formats, analysis)
- annotation for lab data files
- lab data files tracked to samples/patients
- data files upload with provenance
- metadata processing
- lab data collections
- data view & explore
- data access for analysis tools

Lab instrument Data Repository



Laboratory knowledge base

Laboratory Knowledge Base for Instruments, Standards, Data Flow and Data Formats

Use the Data Uploader Form to upload laboratory analysis datasets generated for the Colorectal Cancer OMIC Pilot Study. This form is restricted to authorized members of the CCE OMIC Laboratory Analysis Group.

Learn more about the CCE biological data analysis Laboratories ...

Global Proteomics analysis laboratory in the Purdue University Bindley Bioscience Center.

The instrument-generated datasets and their formats are shown in this LC-MS global proteomics workflow diagram.

Xu Lipidomics analysis laboratory in the Indiana University Simon Cancer Center.

The instrument generated datasets and their formats are shown in this LC-MS lipidomics workflow diagram.

Klaunig Oxidative Stress analysis laboratory in the Indiana University Department of Pharmacology and Toxicology

The instrument generated datasets and their formats are shown in the comet assay workflow diagram and the teac assay workflow diagram.

Raftery Metabolomics analysis laborator in the Department of Chemistry's Raftery Group at Purdue University.

The instrument generated datasets and their formats are shown in the GCxGC-MS metabolomics workflow diagram and the NMR metabolomics workflow diagram.

Regnier Glycoproteomics analysis laboratory in the Purdue University Bindley Bioscience Center

The instrument-generated datasets and their formats are shown in this LC-MS glyco-proteomics workflow diagram.

Teegarden Vitamin D analysis laborator in the Foods and Nutrition Department at Purdue University.

Follow the links to the laboratories for more detailed information about the workflow, instruments, sample preparation protocols, instrument setup methods, generated datasets, formats, conversions, and laboratory analysis.



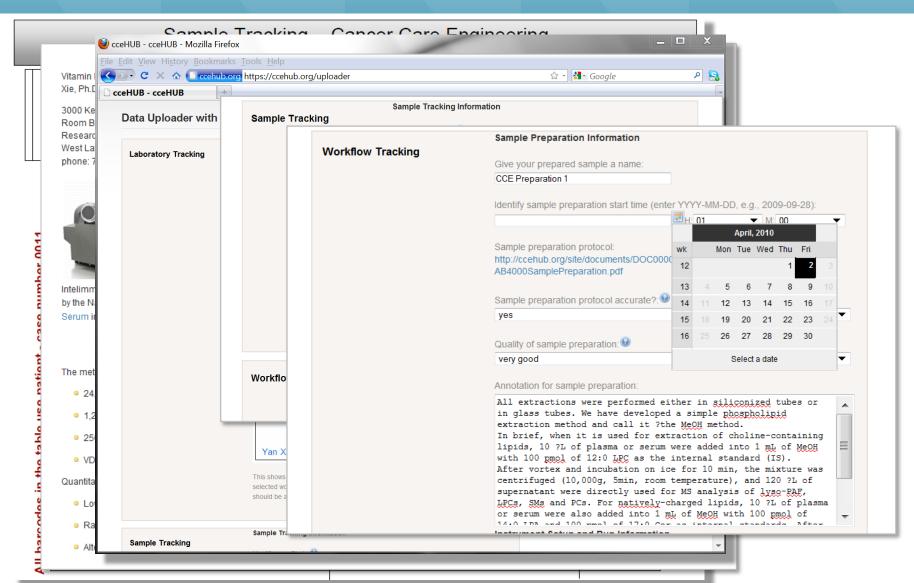
includes only a handful of the original sample group.

Laboratory knowledge base

In the Spotlight Regnier Glycoproteomics Analysis Laborate Process the mass spectral data Posted 07 Feb. 2009 in Series Data Processing and Generated Datasets Data processing will be achieved in two ways. The first will be to transfer the raw MS/MS data containing m/z values and signal intensity to the statistical models and pattern recognition group. The relative amount of peptides will be judged directly from their signal intensity in spectra. The second mode of data processing will be to identify the individual peptides obtained from RPC fractions and their protein parent. This will be done using the MASCOT, Sorcerer or Spectrum Mill search engines. The XCT PLUS ESI ion trap in LC-MS mode generates ".D" files during instrument sample analysis. For peptide quantification, Bruker's CompassXport software is used to convert the XCT PLUS ".D" files to Level 1 LC-MS mzXML files. The mzXML file generated by the XCT PLUS in LC-MS mode has more detailed m/z information than the XCT PLUS run in LC-MS-MS mode. The LC-MS mzXML file can be used as input to the Proteomics Discovery Pipeline for data mining. [You must be logged in to access the Proteomics Discovery Pipeline link.] The mzXML files will also be used by the Cancer Care Engineering statistical modeling group for integrative mathematical modeling. For protein identification, the XCT PLUS LC-MS-MS generated ".D" files are converted to either mzXML Level 2 LC-MS-MS format file used for analysis by Sorcerer software PKL format file generated by DataExtractor, a component of Agilent's Spectrum Mill software MGF, the MASCOT generated format file used for MASCOT database searching. These formats produce peak lists and fragmentation patterns which are matched against databases to identify proteins. Users can access the mzXML and MGF files converted from the D format, as use them as input to other identification software or databases. The PLK file is not accessible after conversion from the D format. In general, the LC-MS and LC-MS-MS phases are separate runs of the XCT PLUS. The LC-MS phase includes hundreds of samples, and the LC-MS-MS phase

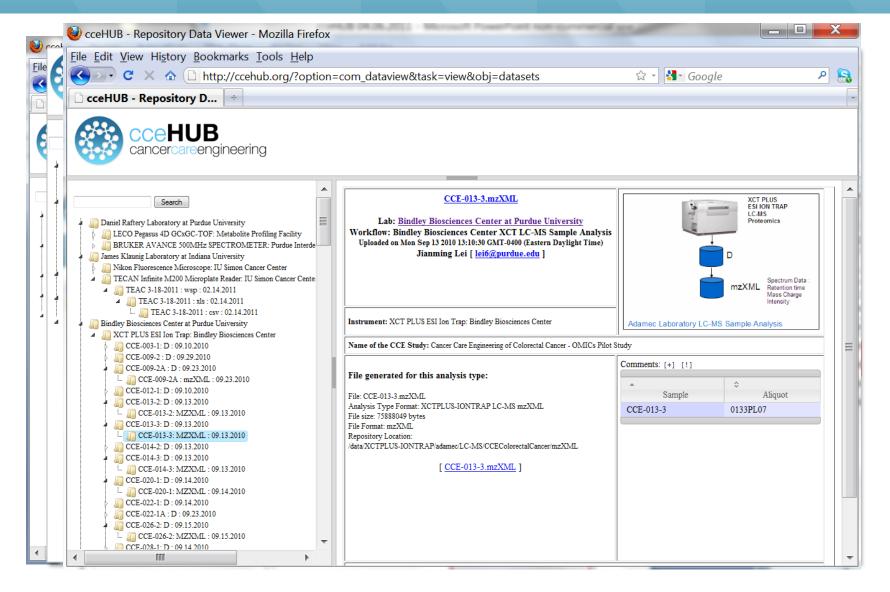


Laboratory data flow, standards, data annotations & upload





Laboratory data repository views





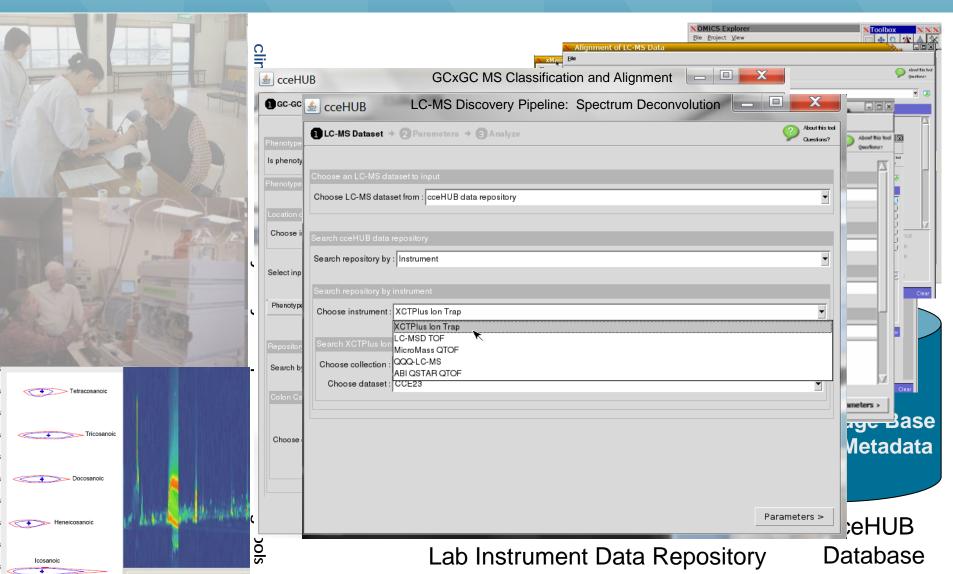
Laboratory data: some stats

Lab	#Samples Analyzed % Total	Files/ Samples Uploaded	Average File Size	Analysis Tools at cceHUB	Using cceHUB tools ?
Bindley Biosciences Global Proteomics	193 73%	193 files 193 samples	80MB	Discovery Pipeline Results Visualize/Compare	500 runs through discovery pipeline
Teegarden Vitamin D	225 85%	4 files 225 samples	< 1MB	Vitamin D-Blood Draw - Clinical Data merge for SAS	Yes, DataView
Raftery Metabolomics GCGC-MS	230 87%	230 files 230 samples	1 GB	Peak classification and alignment GCGC-MS Visual Analytics	
Raftery Metabolomics NMR	110 41 %	1 file 110 samples	< 1MB		
Xu Lipidomics	143 54%	1 file 143 samples	< 1MB	Lipidomics-BloodDraw- Clinical Data merge for SAS	Yes, DataView
Klaunig TEAC analysis	259 98%	1 file 259 samples	< 1MB	TEAC-Blood Draw-Clinical Data merge for SAS	Yes, DataView
Klaunig Comet Assay	101 38%	1 file 101 samples	< 1MB	CometAssay-Blood Draw- Clinical Data merge for SAS	Yes DataView
Klaunig Genotyping Assay				POCRE, MaCH genotype imputation	(used by stat group on their own data)
Regnier Glycoproteomics					



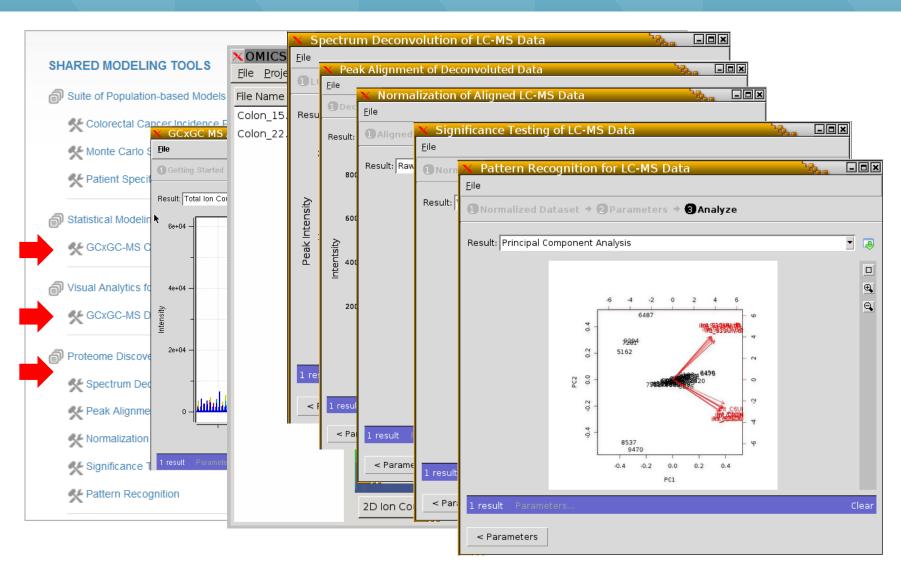
2D Ion Count

Support for modeling and analysis



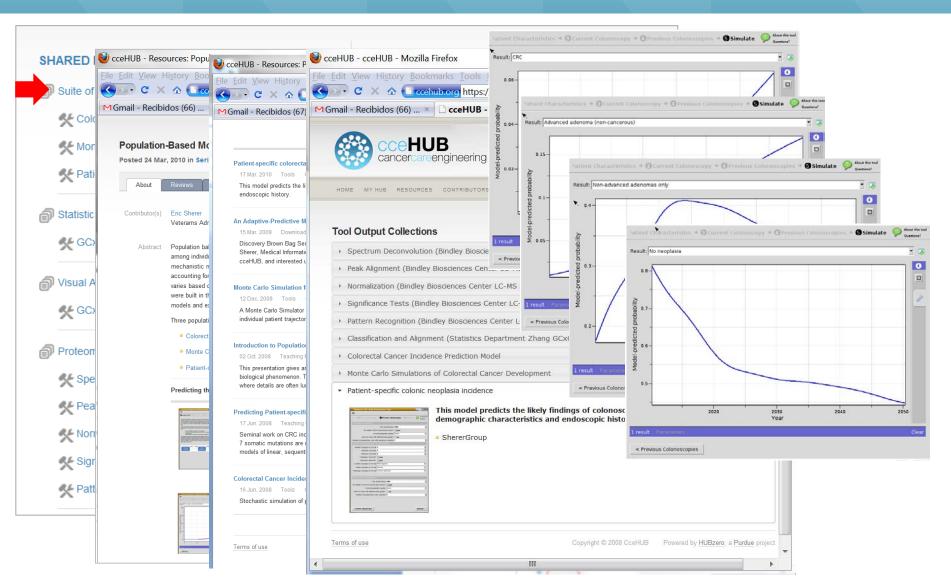


Tools to support data exploration and synthesis



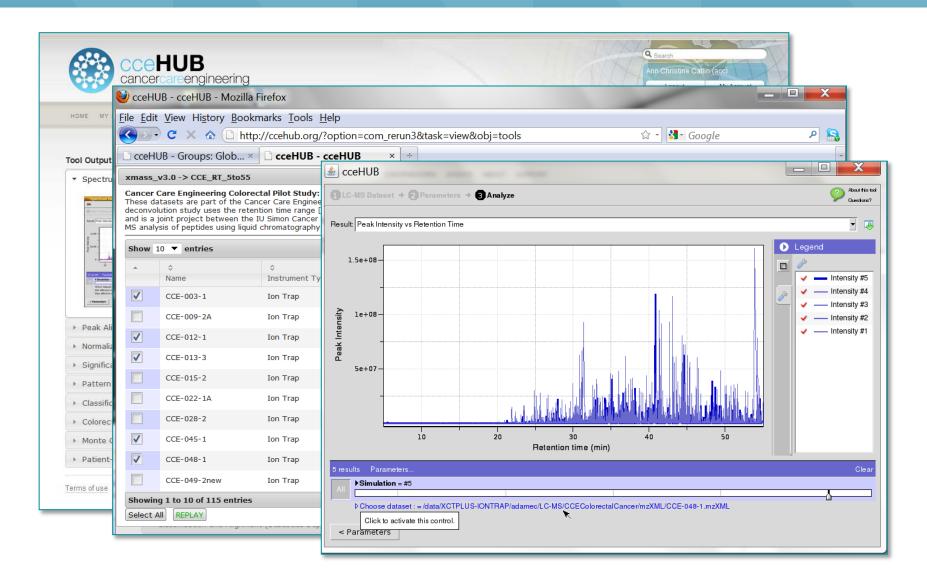


Tools for physician decision support



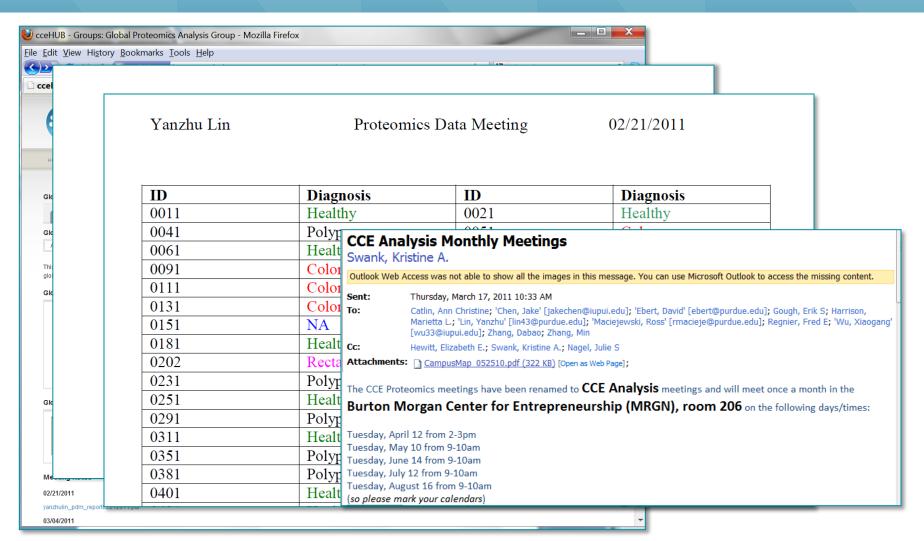


Tool Results Collections and Analysis Browser



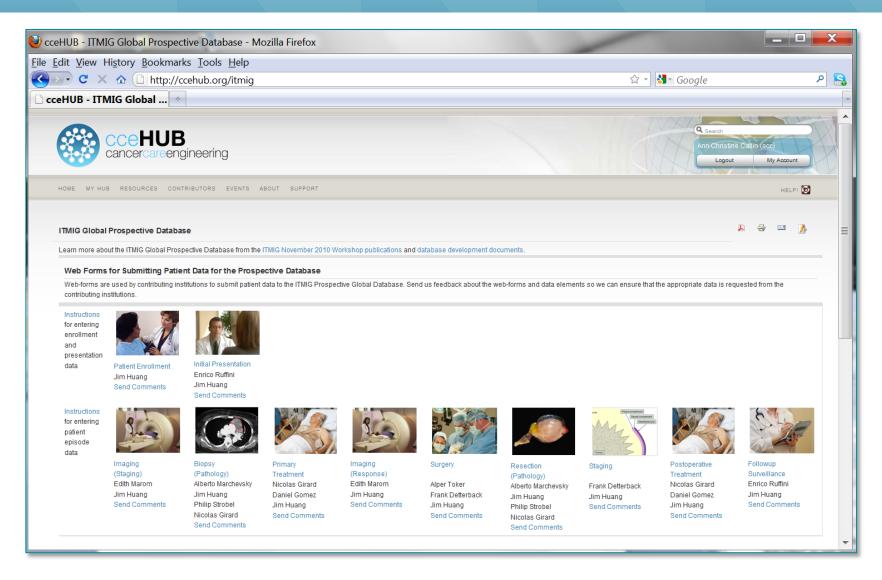


Collaboration for Cancer Care Engineering Research





ITMIG Global Prospective Database just underway ...





Our technology extends to other cancer research workflows

Using the HUB cyber infrastructure and cceHUB data technology to further collaborative Cancer Care Engineering Research

