Bioinformatic approaches to augment study of epithelial-to-mesenchymal transition in lung cancer

Tim N. Beck,1,4 Adaeze J. Chikwem,1,3 Nehal R. Solanki,2,5 and Erica A. Golemis1,3,4

1Developmental Therapeutics Program, Fox Chase Cancer Center, Philadelphia, Pennsylvania; 2Immune Cell Development and Host Defense Program, Fox Chase Cancer Center, Philadelphia, Pennsylvania; 3Temple University School of Medicine, Philadelphia, Pennsylvania; and 4Program in Molecular and Cell Biology and Genetics, Drexel University College of Medicine, Philadelphia, Pennsylvania; and 5Program in Microbiology and Immunology, Drexel University College of Medicine, Philadelphia, Pennsylvania

Submitted 15 May 2014; accepted in final form 4 August 2014

Beck TN, Chikwem AJ, Solanki NR, Golemis EA. Bioinformatic approaches to augment study of epithelial-to-mesenchymal transition in lung cancer. Physiol Genomics 46: 699–724, 2014. First published August 5, 2014; doi:10.1152/physiolgenomics.00062.2014.—Bioinformatic approaches are intended to provide systems level insight into the complex biological processes that underlie serious diseases such as cancer. In this review we describe current bioinformatic resources, and illustrate how they have been used to study a clinically important example: epithelial-to-mesenchymal transition (EMT) in lung cancer. Lung cancer is the leading cause of cancer-related deaths and is often diagnosed at advanced stages, leading to limited therapeutic success. While EMT is essential during development and wound healing, pathological reactivation of this program by cancer cells contributes to metastasis and drug resistance, both major causes of death from lung cancer. Challenges of studying EMT include its transient nature, its molecular and phenotypic heterogeneity, and the complicated networks of rewired signaling cascades. Given the biology of lung cancer and the role of EMT, it is critical to better align the two in order to advance the impact of precision oncology. This task relies heavily on the application of bioinformatic resources. Besides summarizing recent work in this area, we use four EMT-associated genes, TGF-β (TGFβ1), NEDD9/HEF1, β-catenin (CTNNB1) and E-cadherin (CDH1), as exemplars to demonstrate the current capacities and limitations of probing bioinformatic resources to inform hypothesis-driven studies with therapeutic goals.

bioinformatics; cancer; epithelial-to-mesenchymal transition; EMT; NEDD9/HEF1; E-cadherin; SRC; TGF-β; β-catenin; genomics; proteomics; sequencing; precision oncology

The Rising Value of Bioinformatics in Cancer Research

With the increasing use of high-throughput approaches, and after decades of study of molecular mechanisms, it is now apparent that many forms of cancer are associated with a plethora of genetic (113) and epigenetic alterations (246). Despite many recent technological advances (61), and 40 years after cancer DNA sequencing has come into existence (215), the scientific community is still optimizing approaches of data acquisition and interpretation (5, 324, 336). As an interdisciplinary field in which biological, statistical, and computational sciences converge (324), bioinformatics provides essential support for ‘-omics’-driven studies (284). The access to online repositories and web-based algorithms (Table 1) has made it possible to rapidly screen for global and specific changes in gene and protein expression levels as well as genetic alterations (e.g., mutations and translocations) that are unique and sometimes essential to cancer cells (21, 113). For individual investigators, scrutiny of resources containing DNA, protein, and RNA expression data from cell lines, model organisms, and patient samples can enlarge and enhance initial experimental datasets. Subsequent analysis of databases that assign functional annotations to genes [e.g., Unigene (346) and GenBank (26)] and proteins [e.g., UniProt (125), PDB (30), and PFAM (243)], and the association of these elements in signaling networks [KEGG (146), Reactome (65), and others (174)] can assist in the integration of data (Fig. 1).

Precision oncology (21, 101), or individualized cancer treatment guided by -omics data, is dependent on the ability of clinicians to put vast data in a context that allows selection of appropriate therapies. Hence, there is a great need to develop analytic paths that integrate bioinformatic resources to consider the entire system of a tumor. In this review, we describe how bioinformatic approaches can support the study of lung cancer, a type of cancer with great genetic variability (145) and pressing need for better treatments. We particularly address the topic of epithelial-to-mesenchymal transition (EMT), a signaling and morphological shift associated with tumor metastasis.
Table 1. Bioinformatics resources

<table>
<thead>
<tr>
<th>Resource</th>
<th>Description</th>
<th>Location</th>
<th>Ref. No.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genomic Analysis and Visualization</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Babelomics</td>
<td>integrative platform designed to analyze panomics data</td>
<td><a href="http://www.babelomics.org">http://www.babelomics.org</a></td>
<td>193</td>
</tr>
<tr>
<td>GEPAS Analysis Tool</td>
<td>web-based packages for microarray data analysis</td>
<td><a href="http://www.gepas.org">http://www.gepas.org</a></td>
<td>295</td>
</tr>
<tr>
<td>GenePattern</td>
<td>analysis pipeline to process gene expression data, proteomics, SNP analysis, flow cytometry, RNA-seq</td>
<td><a href="http://www.broadinstitute.org/cancer/software/genepattern">http://www.broadinstitute.org/cancer/software/genepattern</a></td>
<td>63, 248</td>
</tr>
<tr>
<td>Gene Set Enrichment Analysis (GSEA)</td>
<td>computational determination of statistically significant gene expression differences between different biological states</td>
<td><a href="http://broad.mit.edu/gsea">http://broad.mit.edu/gsea</a></td>
<td>199, 290</td>
</tr>
<tr>
<td>Genomica software</td>
<td>tool to analyze, integrate and visualize RNA/DNA data</td>
<td><a href="http://genomica.weizmann.ac.il">http://genomica.weizmann.ac.il</a></td>
<td>25, 268</td>
</tr>
<tr>
<td>Geo2R</td>
<td>compare two or more groups in terms of gene expression</td>
<td><a href="http://www.ncbi.nlm.nih.gov/geo/geo2r">http://www.ncbi.nlm.nih.gov/geo/geo2r</a></td>
<td>17</td>
</tr>
<tr>
<td>InSilico DB Genomic Datasets Hub</td>
<td>web-based platform to analyze genomic datasets</td>
<td><a href="https://insilicodb.com/">https://insilicodb.com/</a></td>
<td>63</td>
</tr>
<tr>
<td>Integrative Genomics Viewer</td>
<td>visualization tool for large genomic datasets</td>
<td><a href="http://www.broadinstitute.org/igv/home">http://www.broadinstitute.org/igv/home</a></td>
<td>255, 301</td>
</tr>
<tr>
<td><strong>Integrative Platforms</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bioconductor (R statistical language)</td>
<td>bioinformatics software for the analysis of genomic data</td>
<td><a href="http://www.bioconductor.org/">http://www.bioconductor.org/</a></td>
<td>166, 167</td>
</tr>
<tr>
<td>GenomeSpace</td>
<td>cloud-based platform that accesses bioinformatics tools</td>
<td><a href="http://www.genomespace.org/">http://www.genomespace.org/</a></td>
<td>256</td>
</tr>
<tr>
<td>SAGE/Synapse</td>
<td>platform to share, aggregate and describe research</td>
<td><a href="https://www.synapse.org/">https://www.synapse.org/</a></td>
<td>96</td>
</tr>
<tr>
<td><strong>Genomics and Transcriptomics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COSMIC: Catalogue of Somatic Mutations in Cancer</td>
<td>stores and displays a comprehensive dataset of somatic mutation information and related details</td>
<td><a href="http://cancer.sanger.ac.uk/cancergenome/projects/cosmic">http://cancer.sanger.ac.uk/cancergenome/projects/cosmic</a></td>
<td>89</td>
</tr>
<tr>
<td>Expression Atlas</td>
<td>microarray and RNA-seq data</td>
<td><a href="http://www.ebi.ac.uk/gxa/home">http://www.ebi.ac.uk/gxa/home</a></td>
<td>148, 231</td>
</tr>
<tr>
<td>My Cancer Genome</td>
<td>web-based platform that integrates genomic information, available therapeutics and ongoing clinical trials</td>
<td><a href="http://www.mycancergenome.org/">http://www.mycancergenome.org/</a></td>
<td>291</td>
</tr>
<tr>
<td>Oncomine</td>
<td>genomics information and expression profiles for different cancer types and cancer cell lines</td>
<td><a href="https://www.oncomine.org/resource/login.html">https://www.oncomine.org/resource/login.html</a></td>
<td>252</td>
</tr>
<tr>
<td>The Cancer Genome Atlas (TCGA)</td>
<td>portal to access, search, download and analyse datasets</td>
<td><a href="http://cancergenome.nih.gov">http://cancergenome.nih.gov</a></td>
<td>59, 60</td>
</tr>
<tr>
<td><strong>Cell Lines/Pharmacogenomics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer Cell Line Encyclopedia (CCLE)</td>
<td>comprehensive analysis and detailed characterization</td>
<td><a href="http://www.broadinstitute.org/ccle/home">http://www.broadinstitute.org/ccle/home</a></td>
<td>16</td>
</tr>
<tr>
<td>Comparative Toxicogenomics Database (CTD)</td>
<td>curated database on toxicogenomic interactions</td>
<td><a href="http://ctdbase.org">http://ctdbase.org</a></td>
<td>70</td>
</tr>
<tr>
<td>Pharmacogenomics Knowledgebase (PharmGKB)</td>
<td>curated database focused on pharmacogenomics (i.e., how genetic variation affects drug response)</td>
<td><a href="http://www.pharmgkb.org">http://www.pharmgkb.org</a></td>
<td>331</td>
</tr>
<tr>
<td>The Genomics of Drug Sensitivity in Cancer (GDSC)</td>
<td>cancer cell line-based prediction of responses to anticancer drugs; focuses on compounds and cancer genes</td>
<td><a href="http://www.cancerrxgene.org">http://www.cancerrxgene.org</a></td>
<td>342</td>
</tr>
<tr>
<td><strong>Pathways and Networks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytoscape</td>
<td>software platform to visualize and analyze complex interaction networks</td>
<td><a href="http://cytoscape.org/index.html">http://cytoscape.org/index.html</a></td>
<td>174, 273</td>
</tr>
<tr>
<td>GeneMania</td>
<td>web-based platform used to identify associations of input genes with additional genes as part of a network</td>
<td><a href="http://www.genemania.org">http://www.genemania.org</a></td>
<td>349</td>
</tr>
<tr>
<td>Kyoto Encyclopedia of Genes and Genomes (KEGG)</td>
<td>database resource with focus on high-level functions of biological systems as well as comprehensive pathway maps</td>
<td><a href="http://www.kegg.jp">http://www.kegg.jp</a></td>
<td>147</td>
</tr>
</tbody>
</table>

Continued
and poor response to therapy [Fig. 2; (221)]. We discuss both positive uses of bioinformatic approaches to augment our understanding of these processes and current weaknesses that must be overcome to maximize the value of this form of analysis.

Lung Cancer and EMT

Lung cancer is the leading cause of cancer-related mortality for both men and women worldwide. The majority of lung cancer patients present late in their disease course, with only 15% of cases diagnosed at a localized stage (4a). Nonsmall cell lung cancer (NSCLC, the most common form of lung cancer) has a 5 yr survival rate of ~15%, with metastasis as the primary cause of death (341). Classic treatments for NSCLC include surgery, chemotherapy, and chemoradiation (85). While these treatments are often effective in improving survival, they are also associated with significant toxicities in patients (36, 54, 84).

With the advent of next-generation sequencing, a more comprehensive understanding of the molecular pathogenesis of lung cancer has been achieved (61), offering the hope of more effective and less toxic therapies. NSCLC, typically associated with a history of tobacco use, has been shown to have one of the highest frequencies of mutations among the different cancer types (144, 145, 337). Work of The Cancer Genome Atlas (TCGA) consortium and the International Cancer Genome Consortium has highlighted the many different potential driver and passenger mutations in lung cancer (145, 197). Some of the driver mutations are directly actionable [i.e., activate proteins or pathways that can be targeted therapeutically to block tumor growth (202)]. In the case of NSCLC, valuable protein-targeted therapeutic regimens include drugs inhibiting the epidermal growth factor receptor (EGFR), expressed in 60% of NSCLC and frequently mutated (68, 282). Also, for a subset of NSCLC tumors, drugs such as crizotinib or ceritinib that specifically target EML4-ALK [a fusion protein combining echinoderm microtubule-associated protein-like 4 (EML4) and anaplastic lymphoma kinase (ALK)] have been developed. Some of the available targeted therapeutics are now used as first-line therapeutic agents (45, 56, 93, 276).

To add an additional layer of complexity, lung cancer commonly involves mutations inactivating tumor suppressors, which are generally not therapeutically actionable. TP53 is the most commonly mutated tumor suppressor in lung cancer (135), followed by the serine/threonine kinase LKB1 (270). Despite some improvements in clinical outcomes from targeted therapies, the impact on overall survival rates has thus far been limited (128, 150). This reflects the ability of tumor cells to compensate for inhibition of specific driver proteins (105). In addition, many mutations occurring in NSCLC are currently considered “passenger” mutations. Passenger mutations do not drive tumorigenesis (322); however, some of these mutations can impact disease progression and therapeutic response (190).

One way to target tumor growth more effectively may be to focus on pathological processes specifically associated with

### Table 1.—Continued

<table>
<thead>
<tr>
<th>Resource</th>
<th>Description</th>
<th>Location</th>
<th>Ref. No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathguide</td>
<td>comprehensive list of biological pathway related resources</td>
<td><a href="http://www.pathguide.org">http://www.pathguide.org</a></td>
<td>12</td>
</tr>
<tr>
<td>Reactome</td>
<td>pathway browser to visualize and interact with signaling pathway</td>
<td><a href="http://www.reactome.org">http://www.reactome.org</a></td>
<td>65</td>
</tr>
<tr>
<td>STRING (Search Tool for the Retrieval of Interacting Genes/Proteins)</td>
<td>web-based platform to identify protein-protein interactions</td>
<td><a href="http://string-db.org">http://string-db.org</a></td>
<td>91</td>
</tr>
<tr>
<td>Proteomics Focused</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assisted Model Building</td>
<td>software package to conduct molecular dynamics simulations</td>
<td><a href="http://ambermd.org/">http://ambermd.org/</a></td>
<td>257</td>
</tr>
<tr>
<td>Human Protein Reference Database</td>
<td>information on human proteins</td>
<td><a href="http://www.hprd.org/">http://www.hprd.org/</a></td>
<td>151</td>
</tr>
<tr>
<td>NetworkKin</td>
<td>web-based application to predict kinase-substrate interactions</td>
<td><a href="http://networkin.info/version_2_0/search.php">http://networkin.info/version_2_0/search.php</a></td>
<td>173</td>
</tr>
<tr>
<td>PFAM</td>
<td>database of protein families with focus on functional domains</td>
<td><a href="http://pfam.sanger.ac.uk">http://pfam.sanger.ac.uk</a></td>
<td>243</td>
</tr>
<tr>
<td>PhosphoELM</td>
<td>phosphorylation database with data extracted from literature and phosphoproteomics analyses</td>
<td><a href="http://phospho.elm.eu.org">http://phospho.elm.eu.org</a></td>
<td>75</td>
</tr>
<tr>
<td>PhosphoPoint</td>
<td>comprehensive human interactome of kinases and phosphoproteins</td>
<td><a href="http://kinase.bioinformatics.tw">http://kinase.bioinformatics.tw</a></td>
<td>339</td>
</tr>
<tr>
<td>UniProt</td>
<td>resource for protein sequences and function</td>
<td><a href="http://www.uniprot.org/">http://www.uniprot.org/</a></td>
<td>8</td>
</tr>
<tr>
<td>Molecular Signatures Database (MsigDB)</td>
<td>collection of annotated gene sets that are searchable and that can also be downloaded</td>
<td><a href="http://www.broadinstitute.org/gsea/msigdb/index.jsp">http://www.broadinstitute.org/gsea/msigdb/index.jsp</a></td>
<td>290</td>
</tr>
<tr>
<td>NCBI Homologene</td>
<td>construction of homology groups provides extensive information on transcripts</td>
<td><a href="http://www.ncbi.nlm.nih.gov/homologene">http://www.ncbi.nlm.nih.gov/homologene</a></td>
<td>210</td>
</tr>
<tr>
<td>General Databases</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Physiol Genomics • doi:10.1152/physiolgenomics.00062.2014 • www.physiolgenomics.org
cancer mortality (137, 324) and to disrupt protein interaction hubs central to a given process (137). Metastatic disease is the primary cause of lung cancer-related mortality (341), and for many tumors it is at least partially dependent on EMT (170, 213, 237), although EMT is not the only mechanism of metastasis; for example, some tumors use a different dispersion pattern, by amoeboid movement of individual cells (260). Importantly, tumor cells that have undergone EMT also possess many features of cancer stem cells (184, 229, 263, 265) and are often therapy resistant (6, 40, 50, 67, 194). These features make EMT a process of high interest as a therapeutic target to improve clinical management of lung cancer.

Both intrinsic and extrinsic factors initiate EMT in subsets of primary epithelial tumor cells [evidence suggests single cells or clusters of cells (344)], typically at the tumor margins (329). Although normal and necessary during embryonic development and during wound healing (214), activation of EMT in the malignant neoplastic setting permits tumors to invade surrounding tissue and metastasize to distant locations (37, 64, 340). EMT is not an “all-or-nothing” process, and the roles of EMT in single cells versus in collective cell migration are topics of considerable debate (57, 95, 109, 271). Following initiation, cells first enter a partial EMT state (34, 53, 228, 261), during which cells are quasi-epithelial and quasi-mesenchymal [Fig. 2A; (293)]. Several studies suggest that prior to or during intravasation, the mesenchymal state is assumed, supporting entry of individual tumor cells into blood vessels (251, 311, 312, 344). Alternatively, intravasation has been observed for cell clusters, potentially with or without lead cells that have undergone partial EMT (57, 95, 271). Compatible with the idea of group migration, circulating groups of clustered tumor cells have been found in peripheral blood samples (120, 127). Intravasation is followed by dissemination and eventual extravasation at distant sites (251, 311). For mesenchymal-like cells, once distant sites are reached, mesenchymal-to-epithelial transition [MET; a process also essential during development (176, 214, 298)] is required to allow cells to colonize tissue and form macrometastasis [Fig. 2A; (143, 214, 224, 315)]. The timing of EMT/MET transition is not well established. Several studies provide evidence for both epithelial and mesenchymal circulating tumor cells, as well as cells expressing markers for both phenotypes (9, 169, 344).
EMT is controlled by multiple intrinsic and extrinsic factors [e.g., hypoxia, inflammation and cytokines (142, 170)], which cooperate to regulate gene expression (163), also integrating information about the microenvironmental context of tumor cells [Fig. 2B; (2, 239, 324)]. Intrinsically, there are several commonly used molecular markers for EMT, which include decreased expression of the epithelial marker protein E-cadherin (143, 215, 347); increased expression of the mesenchymal markers N-cadherin (CDH2) and vimentin (VIM) (200); nuclear localization of β-catenin (172); and increased production of the transcription factors (TF) SNAI1 (Snail) (44), SNAI2 (Slug) (323), TWIST1 (340), EF1/ZEB1 (102), SIP1/ZEB2 (258), and/or E47 (66), which control the expression of E-cadherin and other proteins reprogrammed in EMT (163, 297, 318) (Fig. 2). An extended list of EMT-associated molecular markers has been recently proposed (163).

A complicated network of signaling cascades influences regulation of EMT markers and transcription factors. Multiple collaborating pathways, modulated by paracrine and autocrine signaling, activate EMT (263). The main known signaling hubs include the transforming growth factor (TGF)-β superfamily, Wnts, Notch, Sonic hedgehog (Shh), hepatocyte growth factor (HGF), fibroblast growth factor (FGF), epidermal growth factor (EGF), c-Met, and others (103, 188, 203, 215, 238, 299, 308). Scaffolding proteins including NEDD9, p130CAS, DAB2IP, and GRASP/Tamalin also support EMT (11, 41, 216, 308). Scaffolding proteins including NEDD9, p130CAS, DAB2IP, and GRASP/Tamalin also support EMT (11, 41, 216, 308).
FGFR2, RON1, β-catenin, CD44, and others (326). Noncoding microRNAs also control EMT through transcriptional repression of genes governing epithelial identity (162, 348). One of the best-studied families of EMT-associated microRNAs is the miR-200 family, members of which directly target ZEB1 and ZEB2 (Fig. 2B). During EMT these microRNAs are downregulated, resulting in an increase in ZEB1 and ZEB2, thus promoting initiation of the mesenchymal transition (117, 227, 266). In contrast, the EMT-promoting miR-9 binds and inhibits the E-cadherin encoding mRNA and has been shown to induce EMT when overexpressed (181). Additional epigenetic events associated with EMT include reduction of the heterochromatin marker H3K9Me2, coupled with marked increase of the euchromatin marker H3K4Me, at least partially driven by changes in lysine-specific demethylase-1 (Lsd1) expression (189). These epigenetic changes, recently reviewed (293), emphasize the complex nature of EMT and also highlight the importance of acquiring and analyzing large datasets to begin to understand this process.

Currently, tumor-node-metastasis (TNM) staging, a critical clinical tool (186), is not capable of capturing the complexity of cancer in its entirety, as suggested by the heterogeneous survival patterns of patients with identical TNM staging patterns. Thus far, the prognostic potential of EMT markers in lung cancer is limited, and large prospective studies are needed to evaluate and validate the precise clinical value of the EMT markers and signatures mentioned throughout this review. Precision targeting of different cancer cell subpopulations based on EMT status (Fig. 2) is a critical endeavor, and at least four potential opportunities for targeted intervention have been identified: the microenvironment, EMT-associated signaling transduction cascades, mesenchymal cells, and blocking MET at secondary sites (21, 71). A prerequisite for the therapeutic targeting of EMT is precise detection and understanding of the process.

Work by Yu et al. (344) provides a promising model for future clinical studies of EMT. Yu et al. analyzed changes in the epithelial and mesenchymal composition of patient circulating breast tumor cells (CTCs). Interestingly, posttreatment analysis showed that patients with progressive disease while on therapy had an increased number of mesenchymal-like circulation cells compared with patients who responded to therapy (344). Furthermore, mesenchymal-like CTCs had strong TGF-β-induced expression signatures, suggested by the authors to be mediated by TGF-β released from platelets (160, 344). No such study has been completed for lung cancer; however, it has been shown that an EMT-associated secretory phenotype signature that includes TGF-β does have prognostic value in cases of lung cancer (249). The EMT-associated signature strongly correlated with lymph node metastasis as well as tumor stage and grade. Importantly, the signature also predicted survival (249). E-cadherin (discussed below) is another relevant EMT marker that has been shown to be of significance in patients with lung cancer. Expression of E-cadherin has been shown to be associated with favorable outcomes in terms of overall survival in patients with small cell lung cancer (51), and a meta-analysis of 13 studies found that in NSCLC E-cadherin downregulation had an unfavorable impact on survival (332).

It is clear that existing biomarkers must be further validated and additional biomarkers must be identified. The next section describes in detail how bioinformatics has been used to augment the study of EMT in lung cancer. Many of the described resources present critical tools in the effort to study and understand EMT.

**APPLIED BIOINFORMATICS TO ANALYZE EMT IN NSCLC**

In the remainder of this review, we describe the use of orthogonal bioinformatics-based datasets to integrate transcriptome, proteome, and phosphoproteome data. Data integration is used to enrich the understanding of the signaling pathways on which EMT depends and to adequately study this inherently dynamic process (13, 64, 180, 214). Figure 1 in conjunction with Table 1 provides a general workflow and introduces resources useful for the investigation of EMT in lung cancer. As an example of the utility of bioinformatic approaches to extend understanding of EMT, we will focus part of the discussion on a small subset of EMT markers that have been shown to be of relevance in lung cancer (118, 124, 188, 222, 317, 318). These include TGF-β (TGBF1–3) (188), NEDD9/HEF1 (87, 164, 217, 303), E-cadherin (CDH1) (43, 46, 298, 318, 334), and β-catenin (CTNNB1) (299) (Fig. 2). In brief introduction to the group:

TGF-β. The TGF-β ligands (55) regulate pulmonary morphogenesis, wound healing and normal lung homeostasis (19, 188). TGF-β is commonly secreted both by cancer cells as well as tumor stromal cells (88, 187, 209, 230, 278). Thus, TGF-β can play a role in paracrine regulation when it is secreted from stromal cells (281) or via autocrine regulation (134, 292). In the case of paracrine signaling, the heterogeneity of cells within a tumor (106, 176, 192) and the regulatory diversity of TGF-β (acts as both, a tumor growth suppressant or promoter (188, 232)) are factors that can complicate accurate assessment of the phenotypic effects of TGF-β on the tumor. Systematic analysis of single cells from stroma and tumors may be necessary to fully appreciate the connections between the two in terms of TGF-β-induced activities (274).

TGF-β exclusively binds to a hetero-tetrameric transmembrane receptor composed of two type I receptors (TGBFR1, also known as ALK5 and TβRI) and two type II receptors (TGBFR2) and regulates multiple signaling cascades, with the dominant cascade involving SMAD2, SMAD3 and SMAD4 (Fig. 2B; 277). In normal lung cells and during the early stages of tumorigenesis, TGF-β suppresses cell proliferation; paradoxically, in late-stage tumor cells, TGF-β stimulates proliferation, EMT, and invasion (187, 254). Levels of TGF-β expression in NSCLC tumor cells correlate with disease stage (121). Elevated TGF-β detected in the plasma of patients with lung cancer prior to radiotherapy (18, 156, 157), secreted from tumor and stromal cells, correlates with prognosis and also indicates course of disease posttherapy (156). Exposure to exogenous TGF-β can induce EMT in lung cells cultured in vitro (2, 233, 306).

NEDD9. NEDD9 is a noncatalytic scaffolding protein associated with integrin-centered focal complexes, important for cell migration and prometastatic behavior in several types of solid tumors, including those of lung (49, 86, 139, 217, 286). NEDD9 expression in tumors also correlates positively with a poorer overall survival (196). NEDD9 expression is regulated in part by TGF-β signaling, and NEDD9 mediates some TGF-β-induced EMT migratory phenotypes via regulation of effectors including SRC, FAK, and DOCK3/WAVE2 (3, 22,
Reduced E-cadherin expression is correlated with poor prognosis (23, 69, 106, 165, 176). Bioinformatics approaches can add statistical rigor to studies that would otherwise lack sufficient power to reach compelling conclusions, particularly if data from multiple orthogonal approaches are integrated.

The availability of large datasets of lung tumor-associated mutations can be used for these purposes (42, 114, 135), and can samples from model organisms (46) and sets of different established cancer cell lines (1, 16, 100). For example, data from Imielinski et al. reporting the genomic sequencing of 185 lung cancers (135) and TCGA data (42a) are freely accessible through cBioPortal (47, 99) and COSMIC (89) (Table 1). As of May 2014, COSMIC contained information for 26,400 mutated samples for lung cancers and cBioPortal had a combined total of 1,872 lung tumor samples from eight individual datasets. COSMIC has the advantage of having a very intuitive platform layout that effortlessly navigates the user from mutations, to relevant studies on a specified gene, to tissue-specific mutation incidents, to type of mutation (e.g., deletion, insertion), and to drug sensitivity data in the context of a specific mutation. Compared with cBioportal, one downside of COSMIC is that only one term can be searched at a time, and COSMIC only provides information on somatic mutations. cBioportal allows convenient visualization and analysis of large-scale cancer genomics, transcriptomics, and proteomics data (47). cBioportal also allows input of multiple genes and then provides extensive data on co-occurrence of gene alterations, patient survival correlated to alterations of the input gene set, and related changes of 43 phosphoproteins and 117 proteins. Furthermore, in addition to genomics and proteomics information, cBioportal also provides RNA expression levels for some of the searchable datasets.

For any gene of interest, it is now possible to quickly generate an idea of the frequency at which this gene is affected by mutations in tumors and the functional nature of the mutations. Figure 3 shows the workflow to accompany the following example. Typing “Imielinski” into the COSMIC query box rapidly retrieves the 75,708 mutations detected in 185 samples (16,563 genes analyzed per sample) sequences for the study (135). Searching within this extensive list by using the gene symbol for E-cadherin, CDH1, identifies five unique E-cadherin mutations. The specific type of mutation is listed, and each gene is hyperlinked to a page with more information, such as drug sensitivity data for a given mutation, number of samples with a given mutation in the entire COSMIC database (i.e., for all tumor types), sequence and fusion information, and a list of all the studies that identified mutations in a given gene (89). Clicking on a mutant-specific hyperlink reveals affected signaling pathways, which can be extremely useful in terms of hypothesis generation. In the case of E-cadherin, COSMIC identified six affected pathways: apoptosis, cell junction organization, integrin cell surface interaction, signaling in immune system, cell cycle, and Wnt signaling. Further exploration of the EMT-relevant term “cell junction organization,” through a hyperlink to the Reactome webpage (platform for signaling pathway browser; Table 1), provides extensive data on pathways and can also be used to explore the normal expression levels of E-cadherin and related proteins in normal lung tissue (Fig. 4A). These data indicate that in addition to E-cadherin, two cadherin family members (CDH5 and CDH11) are strongly expressed in lung tissue, as are two catenins (CTNNA1 and CTNNA2). Additionally, reduced E-cadherin expression significantly correlated with poor prognosis (23, 69, 106, 165, 176), bioinformatics approaches can add statistical rigor to studies that would otherwise lack sufficient power to reach compelling conclusions, particularly if data from multiple orthogonal approaches are integrated.

The availability of large datasets of lung tumor-associated mutations can be used for these purposes (42, 114, 135), and can samples from model organisms (46) and sets of different established cancer cell lines (1, 16, 100). For example, data from Imielinski et al. reporting the genomic sequencing of 185 lung cancers (135) and TCGA data (42a) are freely accessible through cBioPortal (47, 99) and COSMIC (89) (Table 1). As of May 2014, COSMIC contained information for 26,400 mutated samples for lung cancers and cBioPortal had a combined total of 1,872 lung tumor samples from eight individual datasets. COSMIC has the advantage of having a very intuitive platform layout that effortlessly navigates the user from mutations, to relevant studies on a specified gene, to tissue-specific mutation incidents, to type of mutation (e.g., deletion, insertion), and to drug sensitivity data in the context of a specific mutation. Compared with cBioportal, one downside of COSMIC is that only one term can be searched at a time, and COSMIC only provides information on somatic mutations. cBioportal allows convenient visualization and analysis of large-scale cancer genomics, transcriptomics, and proteomics data (47). cBioportal also allows input of multiple genes and then provides extensive data on co-occurrence of gene alterations, patient survival correlated to alterations of the input gene set, and related changes of 43 phosphoproteins and 117 proteins. Furthermore, in addition to genomics and proteomics information, cBioportal also provides RNA expression levels for some of the searchable datasets.

Bioinformatic Support for Analysis of EMT-relevant Tumor Mutations

Extensive effort has been devoted to the study of individual components of the EMT process (163) and has helped clarify the specific advantages that arise from larger scale studies that incorporate bioinformatics. Such advantages include the identification of biomarker sets of prognostic and therapeutic value (35, 309, 336). These biomarkers also help to validate the appropriateness of models and increase the clinical translatability of experiments (64). Furthermore, due to the transient nature of the EMT process (214) and the typical heterogeneity of tumors (23, 69, 106, 165, 176), bioinformatics approaches can add statistical rigor to studies that would otherwise lack sufficient power to reach compelling conclusions, particularly if data from multiple orthogonal approaches are integrated.
CTNND1) in addition to β-catenin (CTNNB1). This suggests expansion of the query set to evaluate relevance of these proteins to lung cancer pathology and EMT.

Now using the set of highly expressed cadherin:catenin complex proteins found in normal lung tissue (CDH1, CDH5, CDH11, CTNNB1, CTNNA1 and CTNND1) to query cBioPortal highlights that in nearly 50% of lung cancer cases at least one of these genes is mutated, amplified, or deleted or has significantly altered expression (Fig. 4B). Next, it is possible to look at clinical annotations for each sample, which in this case indicating that seven out of eight samples from patients with confirmed, evaluated metastatic disease had alterations in the input proteins. Interestingly, the majority of alterations were transcriptional upregulation, and one striking feature highlighted by this analysis was that none of the six samples had multiple genes altered simultaneously. This suggests the testable hypothesis that an alteration in a single member of this group of cadherin:catenin proteins is sufficient to support changes associated with metastatic disease (Fig. 4B). The provided example highlights how bioinformatics can be used to go from genomics screening data, to interaction networks, to clinically annotated data. This example also emphasizes the value of querying databases with sets of functionally interacting or paralogous genes. It is becoming increasingly clear that mutations and expression-level alterations affecting an important cancer process are often dispersed among members of a pathway regulating that process, rather than invariably clustered on a single pathway member (324). Knowledge of which pathway members are mutated can inform decisions about how to therapeutically target the pathway.

A particular strength of cBioPortal is the convenient testing of hypotheses regarding co-occurrence: i.e., to investigate if gene alterations of interest are more likely to occur in a specific context. For example, Ding et al. (74) identified that the EGFR pathway is the most altered pathway in lung adenocarcinoma (50/188 specimens). Imielinski et al. (135) further reported frequent exonic alterations within EGFR and paralogous kinases (31/183 lung adenocarcinoma tumors). Using cBioPortal to identify mutual exclusion and co-occurrence of gene alterations reveals that EGFR and E-cadherin had a significant co-occurrence (Fig. 5A). Interestingly, Yoo et al. (343) have reported that alterations of E-cadherin/β-catenin predicts poor response to EGFR inhibitors in EGFR-mutation-positive patients treated for lung cancer, noting that EMT is the potential resistance inducer. In this case, biological observations support the potential association between EGFR, E-cadherin, EMT, and therapy resistance; however, this is not always the case and care must be taken to avoid erroneous conclusions.

**Bioinformatic Support for Analysis of EMT-relevant Changes in Gene Expression**

Gene expression patterns have been used to distinguish known types and subtypes of cancer (76, 111, 122), and cancer-profiling studies have been able to identified or reasigned molecular subtypes of cancer (61a, 289). Using gene expression patterns to understand cancer-related processes on the molecular level is being explored. For example, Creighton et al. (64) developed profiles of EMT-associated genes that appeared coordinately across lung tumor specimens collected in TCGA (42a, 130) and a compendium of 11 additional lung adenocarcinoma datasets (representing a total of 1,492 specimens). The complete set was normalized, clustered, and visualized with some of the resources discussed in detail below (Table 1). This approach identified a 16-gene EMT signature [including CDH1, CDH2, fibronectin (FN1), matrix metalloproteinases (MMPs), VIM, and TWIST1; Fig. 2] for use in classifying tumor samples as mesenchymal-like versus epithelial-like (64). The 16-gene EMT signature requires further validation across a larger, independent set of patient samples,
and it is clear that in order to ably capture the prognostic relevance of EMT markers additional steps must be taken. For example, ideally data concerning the microenvironment would be considered alongside each tumor sample, and systematic evaluation of whether the signature is found in large datasets would help ensure that all EMT drivers are identified (64, 163).

The work by Creighton et al. also independently confirmed an earlier study suggesting that low miR-200 family expression levels in lung cancer correlates with expression of EMT markers (110, 117, 228), supporting the ability of high-throughput data mining-based studies to “compete” with gene-targeted low-throughput studies in reaching biological insights.

It is easy to explore the cBioportal database using variations of the 16 gene EMT expression pattern identified by Creighton et al. (64) as representative of an EMT expression profile for lung adenocarcinomas. Combining this search query with a search for NEDD9, β-catenin, E-cadherin, TGF-β, and EGFR confirms low E-cadherin typically correlates with high expression of mesenchymal markers such as CDH2, VIM, SNAIL, and FN1 (Fig. 5A). Additionally, co-occurrence of gene alterations (mutations and expression) were significant for EGFR, TWIST1, and CDH1, suggesting increased EMT; also alterations of the mesenchymal marker CDH2 significantly co-occurred with TGFB1, FN1, and SNAI1. Lastly, TGFB1 alteration had a tendency to co-occur with alterations in the mesenchymal markers FN1 and VIM (Fig. 5A). NEDD9 was altered in 8% of cases; strikingly, the majority of alterations were mRNA upregulation and gene amplification, supporting the role of NEDD9 as an oncoprotein (138, 153, 303). For β-catenin, 12% of cases were altered. Here, the diverse alterations included no clear pattern of mutations, transcriptional up- and downregulation, and proteomic up- and downregulation, suggesting a more nuanced role (Fig. 5B).

Despite the interesting EMT-related connections made possible through the use of databases, a weakness in this approach was also highlighted: a relatively limited number of samples with E-cadherin downregulation are present in the datasets. This may be due to the fact that the biopsies processed to develop gene expression data are not necessarily taken from the periphery of a tumor or metastases, which is where EMT is more likely to be activated (80, 310, 320, 329). Hence, in the case of EMT, the capacity to perform facile and efficient correlative analysis to support hypothesis testing is somewhat limited. The transient nature of EMT and its heterogeneous appearance present a major challenge (163, 312). The dearth of analyzed clinical samples from metastatic sites is likely to be
Carretero et al. (46) used an extensive array of bioinformatics tools to better study EMT in specimens from a mouse model of lung cancer, supplemented with data from patient derived lung cancer samples. This study identified upregulation of EMT-associated genes in metastatic occurrence compared with primary tumors, including \( \gamma \)-catenin signature categories, NEDD9-containing signature categories, as well as signatures containing the NEDD9-interacting proteins SRC and FAK, TGFB1, and CDH1 (22, 217). The resources used in this study included Gene Expression Profile Analysis Suite [GEPAS; (295); recently merged with Babelomics (193)], GenePattern (63, 248), GenomiCA (25, 268), Gene Set Enrichment Analysis [GSEA; (199, 290)], PhosphoPoint (339), and NCBI Homologene (210).

**GEPAS/Babelomics.** GEPAS/Babelomics is a web-based platform that allows the user to normalize, process, and analyze transcriptomics, proteomics, and genomics data and perform integrative analysis and functional profiling as well as data clustering and visualization. Data are easily uploaded, and many input sources are supported, such as the Affymetrix and Agilent platforms, lists of identifiers, lists of annotations, BLAST results, protein-protein interaction sources, and more (193). Normalization can be accomplished by using biconductor, log-transformation, imputation, and replica merging (104, 193). Furthermore, 80 cross-referenced identifiers for genes, proteins, signaling pathways, transcripts, array probes, and

---

**Co-occurrence of alteration of indicated genes**

<table>
<thead>
<tr>
<th>Gene</th>
<th>EGFR</th>
<th>CDH2</th>
<th>CTNNB1</th>
<th>TWIST1</th>
<th>CDH1</th>
<th>FN1</th>
<th>NEDD9</th>
<th>SNAI1</th>
<th>TGFB1</th>
<th>VIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGFR</td>
<td>0.453</td>
<td>0.320418</td>
<td>0.034617</td>
<td>0.008332</td>
<td>0.596466</td>
<td>0.488453</td>
<td>0.087053</td>
<td>0.183988</td>
<td>0.543981</td>
<td></td>
</tr>
<tr>
<td>CDH2</td>
<td>0.21576</td>
<td>0.42305</td>
<td>0.607661</td>
<td>0.506574</td>
<td>0.376093</td>
<td>0.009122</td>
<td>0.001808</td>
<td>0.349268</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTNNB1</td>
<td>0.193046</td>
<td>0.0636252</td>
<td>0.281458</td>
<td>0.316744</td>
<td>0.38058</td>
<td>0.534038</td>
<td>0.279278</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TWIST1</td>
<td>0.33909</td>
<td>0.540763</td>
<td>0.609612</td>
<td>0.122371</td>
<td>0.465962</td>
<td>0.098823</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDH1</td>
<td>0.199629</td>
<td>0.092573</td>
<td>0.597259</td>
<td>0.445148</td>
<td>0.68524</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FN1</td>
<td>0.53722</td>
<td>0.192868</td>
<td>0.02151</td>
<td>0.228838</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEDD9</td>
<td>0.218313</td>
<td>0.707791</td>
<td>0.585364</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNAI1</td>
<td>0.247733</td>
<td>0.543401</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TGFB1</td>
<td>0.013542</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIM</td>
<td>No events recorded for one or both genes</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Fig. 5.** Co-occurrence of alterations in EMT markers in a set of lung cancer samples. A: presented are co-occurrence patterns of alterations in different genes relevant to EMT in lung cancer [cBioPortal; (47, 99)]. B: alteration patterns for the 16-gene EMT signature determined by Creighton et al. (64), in combination with NEDD9, CTNNB1, TGFB1, and EGFR. The vertical row highlighted with dashed lines is a sample with CDH1 downregulation: different mesenchymal markers subsequently upregulated are highlighted with horizontal dashed lines. The black arrows on the left indicate up- or downregulation of mRNA expression levels. The % on the left indicates the number of samples out of 230 total samples that had alterations in the specified individual gene. Dashed lines and black arrows were added manually and are not generated by cBioPortal.

Compensated for in the near future, as the database expands. At present, it can be partially compensated for by incorporating insights from studies using different mouse models addressing EMT (46, 251, 311).

Carretero et al. (46) used an extensive array of bioinformatics tools to better study EMT in specimens from a mouse model of lung cancer, supplemented with data from patient derived lung cancer samples. This study identified upregulation of EMT-associated genes in metastatic occurrence compared with primary tumors, including \( \gamma \)-catenin signature categories, NEDD9-containing signature categories, as well as signatures containing the NEDD9-interacting proteins SRC and FAK, TGFB1, and CDH1 (22, 217). The resources used in this study included Gene Expression Profile Analysis Suite [GEPAS; (295); recently merged with Babelomics (193)], GenePattern (63, 248), GenomiCA (25, 268), Gene Set Enrichment Analysis [GSEA; (199, 290)], PhosphoPoint (339), and NCBI Homologene (210).
functional annotations are available (193, 295). These features make GEPAS/Babelomics well suited to normalize and cluster expression data from different sources (e.g., genotypically different samples, different tissue locations, different disease states, exposed to different pharmacological agents or different concentrations of the same pharmacological agents) to identify gene clusters (i.e., gene signatures) based on common expression patterns (46, 193, 321).

**NCBI Homologene and GenePattern.** For example, Carretero et al. (46) used GEPAS to compare the differential gene expression levels and determine significance of the changes between different mouse cohorts. Given that Carretero et al. used a mouse model to identify genes, it was necessary to use NCBI Homologene, a system designed to allow the query of genes in the context of cross-species homologies [data from 21 completely sequenced eukaryotic genomes are accessible (210)] to compare acquired data to human cases. Subsequently, GenePattern was employed (158, 248) to visualize differently expressed genes. GenePattern is an extremely powerful platform that can perform over 100 different types of analyses. Besides options for interactive adjustment capabilities of visualized output data, supporting advanced gene expression analysis (158), GenePattern allows analysis of proteomic data (98), single nucleotide polymorphism (SNP) analysis, and evaluation of flow cytometry work. GenePattern provides access to the FLAME suite (244) for RNA-seq analysis, profiling DNA methylation patterns (338), exome and whole-genome sequence processing and analysis (165), and more (248). To facilitate the reproduction of experiments by other investigators, GenePattern allows users to establish “analysis pipelines.” In these, standardized sequences of analytical tools applied with consistent parameters to ensure reproducibility (79, 129). Parameters to, for example, normalize data, apply selection thresholds, cluster data, split datasets, and filter output information, are generally selected based on the purpose of the analysis. This is done on a case by case basis (248). Ideally, a saved analysis pipeline is made available alongside published output data and raw data, to allow researchers to reproduce work with precision.

**Genomica.** As an alternative or correlate to GenePattern, Genomica (268) generates a gene expression map based on adjustable parameters, such as up- and downregulation values, false discovery rate (218) correction, and Bonferroni correction (218, 328), and can be used as another practical way of visualizing and characterizing expression data. Genomica is useful for comparing a dataset to predetermined enriched functional gene groups such as gene signatures. Loading two datasets can be done rapidly once this Java-based software has been downloaded. Genomica compares gene sets using hypergeometric distribution based statistical tests and enrichment filters (e.g., $P$ value and minimum/maximum number of enriched genes per set) can be applied. The work by Carretero et al. (46) serves again as a good example. Carretero et al. mined published data on human EMT signatures and used Genomica to compare these signatures to their experimentally derived murine-based EMT data. Importantly, this analysis shows that the signature identified in the murine model correlated with shorter survival and shorter metastasis free survival in human clinical samples (46).

**GSEA.** GSEA, a program similar to Genomica and another alternative for the analysis of gene expression data, is used to compare datasets from samples of different biological states, also with particularly focus on grouping genes (i.e., networks or signatures) that have common biological functions, chromosomal locations, and/or regulation. Comparing sets of genes using GSEA, Subramanian et al. (290) identified common biological pathways between two independent studies of lung cancer patients, for which single-gene analysis provided very limited information.

**GEO2R.** An important concern when relying on databases is whether data collected on different platforms can be legitimately integrated or whether platform-specific artifacts preclude pooling. Addressing this concern, Byers et al. (40) profiled a group of lung cancer cell lines using multiple microarray platforms to establish a gene signature valid across different acquisition platforms (Affymetrix U133A, U133B, Plus2.0 arrays, Illumina WGv2, and Illumina WGv3). Correlation values for the different microarray probes were calculated to identify the most reliable probes. Next, GEO2R was used to compare the different expression datasets. The crux of GEO2R is its ability to effectively and rapidly identify differentially expressed genes across multiple samples (17). It is important for any investigator using GEO2R to realize that this platform accesses and pulls data from the Gene Expression Omnibus database, a public repository, without consideration for experiment type and normalization process. It is up to the investigator to ensure that sample data accessed are indeed comparable. In careful use of this approach, Byers et al. (40) identified a novel EMT-associated EGFR-resistance mechanism involving the receptor tyrosine kinase Axl and its ligand GAS6 and developed a 76-gene EMT signature based on gene expression profiles that can be identified using different acquisition platforms. The work by Byers et al. is promising and clinically relevant. Nevertheless, several potential weaknesses are evident. First, the approach used to determine the 76 genes excluded N-cadherin (CDH2) because it did not meet the established criteria. N-cadherin is a commonly cited mesenchymal marker that was selected by Byers et al. (40) to define the mesenchymal category, yet it still failed to meet the criteria necessary to be included in the signature. Second, the 76 genes were selected based on 54 NSCLC cell lines, an approach that fails to appreciate the input of the microenvironment. This may explain why Byers et al. did not find the EMT signature to have any prognostic value in terms of disease control or progression-free survival. Independent validation of the 76-gene signature across a larger sample set (Byers et al. evaluated only 139 samples) would help establish the potential of this signature (165).

**Bioinformatic Support for Analysis of EMT-relevant Protein Signatures**

A current frontier of analysis of cancer biology is in the area of proteomics. The issues discussed above limiting the effectiveness of gene expression databases informing EMT, notably, biopsies with limited representation of specimens from tumor margins or metastases, and the transient nature of the EMT process, at present also apply to protein-based data. Nevertheless, some work is beginning to suggest approaches that can be further developed, as technologies and databases improve. For example, Carretero et al. (46) identified hyperphosphorylation of a group of established SRC substrates as a signature present
in primary lung tumors and metastases, compatible with the upregulation of NEDD9 and its partner SRC in tumors that are undergoing EMT. First, phosphoproteins were isolated from murine tumors as well as human cancer cells using immobilized antibodies, followed by liquid chromatography-tandem mass spectrometry analysis to establish phosphorylation profiles. To find putative upstream kinases for the identified phosphoproteins, four independent databases were used to maximally capture information on kinase target site specificity: PhosphoELM (75), HPRD (151), Swissprot/UniProt (8), and NetworkKIN [Table 1; (46, 173)].

Based on the hyperphosphorylation data as well as identified gene expression patterns, Carretero et al. (46) hypothesized that the triple combination of dasatinib (Src inhibitor), AZD2644 (MEK1/2 inhibitor), and BEZ135 (PI3K-mTOR inhibitor) would be a viable treatment approach. The triple combination caused significant tumor regression in vivo and was superior to dasatinib alone or dual drug combinations. As SRC inhibitors specifically, and targeted inhibitors in general, tend to be most successful as part of combination therapies, this class of approach has obvious value (38, 154, 242, 247).

An alternative approach for the assessment of proteomic change is the use of reverse-phase protein arrays [RPPA; (302)]. RPPAs use antibody panels to query over 100 cancer-relevant proteins for expression or activity (as reflected by antibody-detectable phosphorylation sites). Byers et al. (40) integrated RPPA-derived proteomic data for epithelial and mesenchymal cells by hierarchical clustering, also incorporating, for comparative measures, analysis of different microarray platforms to analyze EMT signatures (discussed above). The proteomic analysis showed that E-cadherin was the most significantly different protein between the epithelial (highly expressed E-cadherin) and the mesenchymal group (low expression of E-cadherin). Importantly, E-cadherin mRNA expression levels varied significantly for the different E-cadherin probes, in terms of correlation to the protein levels, highlighting the need to consider a signature of genes rather than a single gene, due to acquisition associated variation, for transcriptional analyses and preferably integrating proteomic data. Byers et al. (40) then tested several drugs on mesenchymal and epithelial cell lines, categorized according to their data analysis, to establish sensitivity and resistance patterns. Cell lines classified as more mesenchymal in characteristics were more resistant to EGFR and PI3K/Akt inhibitors, confirming the validity of the informatics predictors. Importantly, bioinformatics enabled application of the same analysis pipeline used for cell lines analysis to analyze lung cancer tissue samples from patients, and the experimentally determined EMT signature was concluded to potentially serve as a marker of erlotinib activity in some patients (40). This example thus simultaneously emphasizes the value of informatics approaches in suggesting clinically useful predictors but at present also indicates the need to use multiple probe sets to evaluate gene expression data as well as proteomics data to identify predictive patterns.

Building Simple Local Relational Networks for Proteins of Interest

There are other ways web-based tools can quickly and conveniently support studies of genes nominated as of interest for EMT in lung cancer (or any other purpose). For example, we used cBioPortal (47, 99) and two additional platforms [STRING (91) and GeneMania (349)] to investigate proteins and genes functionally and physically interacting with NEDD9, with emphasis on SRC (Fig. 3). All three programs are web-based, free, and fairly intuitive to use on a basic level. Figure 6A highlights one of the available cBioPortal features. Shown is the analysis of human lung adenocarcinoma datasets that examines different protein expression levels of SRC and phosphorylated SRC using SRC and NEDD9 as the query input. This was done to test the hypothesis that the expression of these proteins was coordinately regulated. The analysis showed that genomic or transcriptional alterations (predominantly amplification and mRNA upregulation) of NEDD9 in lung cancer correlates with increased levels of p-SRC (pY527), the most significantly changed protein out of 47 tested phosphoproteins, and total levels of SRC (P = 0.000212 and P = 0.019 respectively; Fig. 6A), supporting one potential mechanism by which NEDD9 contributes to EMT. cBioPortal provides an extensive list of altered protein expression levels, with associated P values; additionally, data on mutations, coexpression, survival data and more, for genes of interest are also provided.

Next, we used STRING (91) and GeneMania (349) to investigate how SRC and NEDD9 relate to each other in the context of a protein interaction networks (Fig. 6, B and C). STRING and GeneMania access multiple databases reporting protein-protein interactions and generate simple networks around query proteins within seconds. Networks centered on NEDD9 and SRC show close links between the two proteins (Fig. 6, B and C) and also indicate interactions to other proteins of relevance, providing links to the information used to construct the databases. Hence, this form of analysis can help support and organize literature searches. Notably, the networks generated by the two programs are not identical. For example, STRING finds evidence for a direct link between NEDD9 and SRC, whereas GeneMania does not. As always, the individual investigator must use judgment and confirm suggested results in the literature or by experiment. Liu et al. (174) provide a detailed instructional guide to generating both simple and more complex interaction networks, and they provide extensive information regarding advanced functions of STRING and GeneMania, as well as information on many other available programs to extend this form of analysis.

Current Limitations of Bioinformatics

Because of the great investments being made in the generation of high-throughput data and the increasing dependence of clinical management of cancer patients on genomic information (see below), bioinformatics is inevitably assuming a greater role across many disciplines. Within the next decade, continued improvements in resources will help to organize the complexity of the EMT process. However, it is important to realize some of the limitations of bioinformatics (Fig. 1). First, the inconsistencies introduced by different acquisition platforms (40) as well as those introduced during the bioinformatics analysis itself (79, 129) have been mentioned to some extent above. Working closely with committed bioinformaticians and statisticians best minimizes both of these aspects. At the same time, bioinformatics can be used successfully to expose inconsistencies and to cross-compare data that theore-
ically should correlate. For example, Haibe-Kains et al. (119a) found that two large pharmacogenomic studies (16, 100) presented expression profiles in cell lines with remarkable correlation between the two studies. Unfortunately, the drug response data presented in the same two studies were highly discordant. The most likely explanation for the inconsistent drug response data is the use of slightly different assays and experimental protocols, unfortunately a common issue in high-throughput studies (24). Similar problems encountered for gene expression data were at least partially addressed with the establishment of strict standards (207), which would be desirable for future pharmacogenomic efforts.

Fig. 6. Exploring changes in protein expression and building simple local protein-protein interaction (PPI) networks. A: expression levels of phospho-SRC and total SRC are shown in a comparison of samples with no alterations for NEDD9 and SRC (Unaltered) to samples with at least one alteration in NEDD9 or SRC (Altered). RPPA data were accessed with cBioPortal; RPPA score (on y-axis) refers to the z-score for each sample based on average protein abundance (47, 99). A significant difference exists between the 2 categories (Altered vs. Unaltered) for both phospho-SRC and total SRC. STRING (B) and GeneMania (C) were used to build PPI networks to highlight the relationship between NEDD9 and SRC. For the network generated from STRING (91), only data with medium or higher confidence scores [scores are computed by combining the probabilities for a given interaction based on evidence from different sources and correcting for the probability of random observations (325)] obtained by limited methods (coexpression, gene fusion, and experiments) were included to generate the displayed interactions (blue lines). Only physical interactions (pink lines) are shown for the PPI network generated with GeneMania (349). Protein nodes were manually arranged for both PPI networks.

Discordance can also be observed in the simple interaction network presented for NEDD9 (Fig. 6, B and C), emphasizing the selective nature of algorithms using different metrics to probe similar resources. Where possible, the approach taken by GenePattern (noted above; Table 1) in establishing standard analytical pipelines will help ensure reproducibility (248). The problem of limited exchange among scientists in terms of computational analyses of published data remains and hinders standardization and consistency. Considering genes as groups rather than on an individual basis may be most appropriate, for example, by looking at signaling pathways and biological processes such as EMT as functional units (29, 290, 324). Grouping individual data points may also help address another limitation of bioinformatics in terms of its clinical relevance: its propensity to overwhelm clinicians and patients (28). One promising emergence prompted by the need for facilitated collaboration among scientists and institutions in terms of data sharing is SAGE/Synapse (96). In very broad terms, the SAGE/Synapse platform can help scientists address several fundamental hindrances, such as finding data, understanding workflow, analyzing data, and establishing viable and vibrant collaborations (Table 1). It is critical that information is presented in a comprehensible manner, in a way that saves time and directs decision-making, instead of simply adding more items to a decision-making tree, items that may or may not be significant.

**BIOINFORMATICS IN SUPPORT OF PRECISION ONCOLOGY**

Precision oncology integrates cancer genomics, metabolomics, proteomics, and transcriptomics technologies to inform the treatment strategies for individual patients (280). The discovery of the Philadelphia chromosome, a translocation leading to a constitutively active BCR-ABL tyrosine kinase first described in chronic myelogenous leukemia (CML) (220), and the subsequent development of the first compound for targeted cancer therapy, imatinib (78), which binds to and inhibits the activated BCR-ABL oncogene, can be considered the first major and quite revolutionary breakthrough of precision oncology. This treatment approach has had a dramatic impact on patient survival in some cases (77), and the success of imatinib initiated the progressive introduction of a new

Physiol Genomics • doi:10.1152/physiolgenomics.00062.2014 • www.physiolgenomics.org
category of treatment, one not exclusively based on location and histology, but also considerate of specific mechanistic aspects driving pathological progression.

To date, precision oncology has sought to identify mutations in driver oncogenes and to suggest ways to therapeutically target these drivers (324). Failure to understand the feedback circuits between signaling pathways currently limits the effectiveness of targeting mutated drivers with matching drugs, contributing to the problem of unexplained de novo and acquired resistances (112). As precision medicine is evolving, it has become evident that precision medicine requires precision data acquisition, precision data analysis, precision data interpretation, and, last but not least, precision communication between the medical team and the patients. Appropriate bioinformatics helps address many of these requirements (96, 101, 280, 291). Given the linkages between EMT and drug resistance noted above, incorporation of datasets addressing this process will inevitably assume a larger role in therapeutic decision making in coming years.

Matching Therapy to Patient: Targeting EMT

At present, few therapeutic agents directly target proteins involved in EMT for either lung cancer or other cancers. Instead, driver oncogenes targeted in specific subsets of lung cancer patients include EGFR (135, 198) or EML4-ALK (159, 276). When matched to the appropriate patients, these have remarkable response rates. Not surprisingly, an ever increasing number of cancer-focused clinical trials require in-depth analysis of patient biopsy samples to determine drug response markers (20), a trend that is likely to continue. The key prerequisites for clinical translation of an individualized treatment approach into a successful early-phase trial still present a formidable challenge. First, reproducible and uniform characterization of the genomes of patients’ tumors by state-of-the-art technologies must be achieved (using samples from multiple tumor sites). Second, the genomics data must be filtered through a knowledge base of existing and emerging anticancer drugs to select the optimal treatment regimen. Third, an annotated list must be presented to the treating oncologist to inform clinical decisions (90). However, multiple challenges must be addressed before this can be done. Tumor samples are substantially heterogeneous genetically, with a mixture of malignant and nonmalignant cells, as well as being composed of different clones and subclones within individual tumors (195, 294). Even though many of critical points remain to be addressed, it is encouraging that several clinical trials are already investigating the possibility of using molecular profiling to guide therapeutic interventions (152, 168, 313, 314).

When considering targeting EMT in the context of precision medicine, there are additional issues. First, as noted above, the EMT program is not active in all tumor cells: it is most active at tumor margins and in metastasizing or recently metastasized cells. It is harder to identify the correct “target” population. Second, although some driver oncogenes and tumor types are significantly associated with a higher incidence of EMT (i.e., melanoma is significantly more metastatic than prostate cancer), the EMT program itself is more likely to be dynamically regulated at the epigenetic level, again separating the phenomenon from readily targetable upstream drivers. Third, one theoretical objection to using drugs to target EMT arises from the pragmatic handling of patients receiving initial diagnoses in the clinic. Before EMT has occurred, in situ tumors are excised surgically. Conversely, if tumors have already metastasized, common in lung cancer diagnosis, one could argue that it is too late: EMT has already occurred. Hence, this would limit the primary use of EMT-targeting agents to early invasive tumors, which constitute a small subset of the whole.

An answer to these objections lies in consideration of the genes identified by bioinformatics as part of the EMT signature and the growing recognition that there is significant overlap between features of cells that have undergone EMT with those in a drug-resistant “stem cell-like” compartment (184, 237, 285). Targeting intermediate elements of EMT cascades (Fig. 2B), such as SRC (97, 108, 287) or RhoA (32, 205, 272), may be broadly useful in reducing survival of cells treated with cytotoxic agents by removing them from stem cell-like differentiation states. Additionally, and again returning to clinical practice, many lung tumors are diagnosed in elderly or frail patients, for whom surgical resection of a primary tumor poses significant risks or who cannot tolerate the side effects of chemo- or radiotherapy. In some of these patients, watchful waiting is advised (155, 267). If a well-tolerated EMT inhibitor were available for use during this period of watchful waiting, this might significantly reduce the later appearance of metastases in this vulnerable patient group. Furthermore, considering that 90% of human cancer deaths are due to metastases and much of the disease’s incurability is due to metastases (46, 48), inhibition of EMT to limit the deadliest aspect of cancer would have a tremendous impact.

Based on bioinformatics approaches, what EMT proteins might make good targets? The problem is not just identifying potential EMT-related proteins that may make good targets based on biological roles, but also the feasibility of targeting such proteins pharmacologically. Here, too, bioinformatics resources can provide data that assist in prioritizing targets and identifying small molecules that may interact with a given target. One of the most comprehensive resources available to investigate “drug-event” datasets is the Comparative Toxicogenomics Database [CTD; (70)]. CTD covers 23,200,343 toxicogenomic relationships, which includes 969,979 curated chemical-gene interactions and 1,618,527 chemical-disease associations (190,935 of them are curated). Different categories can be used to narrow searches, such as Gene Ontology (GO), chemicals, disease, gene, organism, pathways, or references. Selecting the GO category and using “EMT” as the input term provides a brief description of EMT and makes available a tab with established genes involved in EMT. CTNNB1 is one of the genes listed; clicking on the accompanying hyperlink opens a site with the top 10 interacting chemicals, one of which is indomethacin, a nonsteroidal anti-inflammatory drug. CTD further provides the references for each indomethacin-CTNNB1 interaction, with a one-line summary statement of the findings for each reference. In this specific case, the references suggest that indomethacin may result in decreased CTNNB1-associated pathogenesis. Thus, studying the impact of indomethacin in the context of EMT specifically could be of interest. A similar approach using “CDH1” as the search term indicates that afimoxifene results in increased expression of CDH1. Considering the interaction between CDH1 and CTNNB1 (Fig. 2B), it would be reasonable to hypothesize that combining indomethacin with afimoxifene could limit EMT;
however, clearly extensive biological studies would be necessary to further this idea (Fig. 1), but the general notion of how CTD can be used to think about novel drug combinations based on known biology should not be dismissed. Additional useful resources for gene-chemical interactions are listed by Pathguide, which provides a list of 547 (as of May 2014) biological pathway and molecular interaction-related resources (12).

Predicting Therapeutic Resistance: a Role for EMT?

Unfortunately, at present almost all patients receiving targeted therapy experience relapse (161). Study of relapse mechanisms suggests some relevance to induction of EMT. For example, in EGFR-mutant lung cancer, amplification of c-MET (83, 316), mutations in PIK3CA (269), or mutations in BRAF (223, 241), which allow bypass activation of critical downstream signaling, provide resistance to EGFR-targeting drugs. c-MET activation is a particularly interesting resistance mechanism, as it not only induces EMT but also involves several of the aforementioned proteins to do so (107). Xi et al. (333) showed that c-MET activation promotes increased phosphorylation of the NEDD9-associated protein SRC (3, 287), which reduces the interaction of β-catenin with E-cadherin, and instead leads to β-catenin phosphorylation, nuclear translocation and EMT induction. Lung tumors have also been shown to form resistance by transforming to small-cell lung cancer, as well as EMT (52, 269). The molecular mechanisms of EMT-associated resistance to EGFR inhibitors, observed both in patients (269) and cell lines (62, 92), is not yet fully understood; the broader EMT and EGFR-inhibitor-associated expression changes described above, obtained through bioinformatics means, should contribute to understanding these processes (40, 46).

Pharmacogenomics, noted briefly above, is an emerging field that attempts to understand how genomic information can be used to predict therapeutic responses to pharmacological agents (123). Specific mutations and variations can serve as markers used to identify subpopulations that are either more or less likely to benefit from certain therapeutic agents, succumb to adverse drug reactions, or require variations in drug dosing. Pharmacogenomic DNA markers, aside from identifying viable tumor targets, are also critical to fully appreciate drug absorption, distribution, metabolism, and elimination (191, 330). The best-known pharmacogenomics resource currently available is PharmGKB (300, 331), although other databases do exist (283). PharmGKB covers clinical information (e.g., dosing guidelines, drug labels, clinical annotations, and associated publications) and genotype-drug-response relationships. In the case of lung cancer, the most relevant gene covered by PharmGKB is EGFR, described here as an example of the potential use of this resource. Hodoglugil et al. (126) used PharmGKB to give a pharmacogenomics overview of EGFR in the treatment of NSCLC. First, the authors took advantage of the pathway descriptions provided by PharmGKB for different targets, which, in the case of EGFR, breaks down available pharmaceuticals into two categories: tyrosine kinase inhibitors [bind the EGFR kinase domain intracellularly (175)] and monoclonal antibodies [block EGFR ligands from binding; (175)]. Next, PharmGKB was used to pull up a table of known pharmacogenomic EGFR variants. The authors then summarized the information on specific variant rs11568315, the occurrence of which is linked to a better clinical response to the tyrosine kinase inhibitor gefitinib (307). The benefit of quickly and effectively accessing pharmacogenomics data is clearly of immense value to both researchers and clinicians; however, more data are needed as well as expansion to other genes linked to therapeutic resistance, such as those in EMT signaling cascades.

Specific pharmacological inhibition of EMT poses plentiful challenges. One potential approach to blocking or reversing EMT is by increasing E-cadherin expression (178), which could be achieved by inhibiting negative regulators of surface E-cadherin. For example, SRC inhibition significantly increases E-cadherin in vitro, suggesting the use of SRC-inhibitory agents such as dasatinib for this purpose in vivo (206). However, it is unlikely that targeting a single protein will limit EMT in a clinical setting. Targeting multiple EMT relevant and E-cadherin-associated proteins simultaneously may be a more promising approach. For example, using the histone deacetylase inhibitor vorinustat, which has been shown to prevent TGFβ1-induced decrease of E-cadherin (133), in combination with an SRC inhibitor would be a hypothesis-driven, rational drug combination study focused on limiting EMT. An alternative approach to increase E-cadherin may be treatment with the proteasome inhibitor marizomib, a drug observed to increase E-cadherin and inhibit Snail, leading to decreased tumor invasion and migration (15). Yet another potentially promising approach is targeting signaling proteins downstream of membrane-bound receptors, such as TGF-β and EGFR, which induce EMT (115, 179). First preclinical studies have shown that inhibiting MEK or ERK1/2 does indeed induce epithelial markers and limit invasiveness (39). Clearly, most of these proteins also have functions beyond a role in EMT: clinical gain may come from parallel inhibition of multiple processes that support malignancy.

Finally, tumor heterogeneity is inherently linked to EMT (183, 184, 251, 329). Heterogeneity within individual tumors (clonal and subclonal mutations; microenvironment-specified epigenetic variation) provides a potential explanation for both innate and acquired drug resistance. Targeted therapy against one driver mutation could select for the enrichment and clonal expansion of a small number of cells resistant to that agent (42, 116). An intriguing mathematical model (7) has recently been applied by Bozic et al. (38) to explore the possibility of anticipating cancer evolution and treatment failure based on patient-derived data. Upon application to a set of patients with varying tumor burden, the model suggested that monotherapy with a targeted therapeutic will result in resistance, whereas simultaneous dual therapy may provide long-term disease control; however, for patients with large disease burden, triple therapy was suggested. In this case, bioinformatics provides a framework for potential advantages of reconsidering the design of clinical trials for targeted therapies, and it provides expectations for targeted therapy in general (Fig. 1). Future efforts will also need to address the heterogeneous nature of tumor microenvironments in driving drug resistance and EMT. At present, fewer computational models take into account relevant biological factors, such as the abundance of growth factors derived from tumor-adjacent stroma (Wnt, TGF-β, FGF, EGF, and others) and the local inflammatory environment (14, 94, 141, 160, 245), even though these are highly relevant: ulti-

Physiol Genomics • doi:10.1152/physiolgenomics.00062.2014 • www.physiolgenomics.org
mately, bioinformatic models will need to factor in these important variables.

CONCLUSION AND FUTURE DIRECTIONS

Although transient, EMT is a critical aspect of cancer biology and over the past decade significant headway has been made in terms of understanding this process. Much is still to be learned. From a therapeutic standpoint, as targeted drugs impose varying selective pressures, resistance emerges through dynamic events affecting proteins integral to the EMT program and other signaling pathways (81). Adding inter- and intratumor as well as interpatient heterogeneity (23, 69, 106) into the mix, and it becomes clear that in the absence of a systems-level understanding, the individual oncologist will not typically be able to therapeutically exploit data from ever more detailed molecular characterization of tumors. While basic investigations remain necessary to understand the core biological processes of EMT and metastasis, current efforts to establish comprehensive, interoperable, and easily accessible databases are necessary to drive progress (Fig. 1). In the past, there have been calls to make training in basic computational and informatic approaches a component of graduate education in biology and medical school: the need has never been more apparent.

Precision oncology is the future of cancer treatment, with many clinical trials already investigating the feasibility and impact of this approach (152, 168, 313, 314). The promise is tremendous, but the challenges remain significant. Considering that targeted therapy still benefits only 10–20% of patients for only 6–12 mo (128), the road ahead will not be easy. It is also clear that monotheraputic therapy fails rapidly and even dual-targeted therapy is unlikely to provide long-term relief or even be curative (38). Though speculative, we think that solutions are on the horizon. The challenge of target identification and disease evolution as well as disease heterogeneity may be addressed via serial liquid biopsies and analysis of circulating tumor cells and circulating tumor DNA (4, 31, 33, 72, 73, 344). In combination with the advent of single cell sequencing (82, 208, 240), this approach should make it possible to accurately define and track an individual’s pathology. Improved intravital microscopy will allow better investigation of dynamic cancer-related processes such as EMT and will help fill in some of the missing links (163, 234, 262). The expansion of current databases will provide the necessary reference resource to help predict the behavior of an individual’s disease. For these efforts to yield fruit, standardization of data acquisition processes, data normalization and analysis procedures and globe-wide access to the data are essential. Finally, new drug development should benefit from the wealth of data, and systems such as personalized “organs on a chip” (131, 132) may ultimately provide the tools needed by the industry to truly personalize oncology.

GRANTS

The authors were supported by National Cancer Institute Grants U54 CA-149147, R21 CA-181287, R01 CA-63366, and P50 CA-083638 (to E. A. Golemis), F30 CA-180607 (to T. N. Beck) and Core Grant CA-06927 (to Fox Chase Cancer Center).

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS


REFERENCES

11. Attar MA, Santy LC. The scaffolding protein GRASP/Tamalin directly binds to Dock180 as well as to cytoskeletons facilitating GTPase crosstalk in epithelial cell migration. BMC Cell Biol 14: 2013.


Fusaro VA, Mani DR, Mesirov JP, Carr SA. Expression profiling of small cellular cytoplas.


Review

718  APPL YING BIOINFORMATICS TO STUDY EMT IN LUNG CANCER


Downloaded from http://physiolgenomics.physiology.org/ on August 11, 2017

Physiol Genomics • doi:10.1152/physiolgenomics.00062.2014 • www.physiolgenomics.org
Applying bioinformatics to study EMT in lung cancer

719


Review

APPLYING BIOINFORMATICS TO STUDY EMT IN LUNG CANCER


