Cell

Identification of Causal Genetic Drivers of Human Disease through Systems-Level Analysis of Regulatory Networks

James C. Chen,[1,2,3] Mariano J. Alvarez,[1,2] Flaminia Talos,[3] Harshil Dhruv,[5] Gabrielle E. Rieckhof,[1] Archana Iyer,[1] Kristin L. Diefes,[6] Kenneth Aldape,[7] Michael Berens,[5] Michael M. Shen,[1,3,4,8,10] and Andrea Califano[1,2,9,10,*]
[1]Department of Systems Biology, Columbia University, 1130 Saint Nicholas Avenue, New York, NY 10032, USA
[2]Center for Computational Biology and Bioinformatics, Columbia University, 1130 Saint Nicholas Avenue, New York, NY 10032, USA
[3]Department of Genetics and Development, Columbia University, 701 West 168th Street, New York, NY 10032, USA
[4]Department of Medicine, Columbia University, 630 West 168th Street, New York, NY 10032, USA
[5]Cancer & Cell Biology Division, TGen, 445N 5th Street, Phoenix, AZ 85004, USA
[6]Department of Pathology, M.D. Anderson Cancer Center, University of Texas, 1515 Holcombe Boulevard, Houston, TX 77030, USA
[7]Adult Brain Tumor Centre, Ontario Cancer Institute, University of Toronto, 610 University Avenue, Toronto, ON M5G 2M9, Canada
[8]Department of Urology, Herbert Irving Pavilion, Columbia University, 161 Fort Washington Avenue, New York, NY 10032, USA
[9]Department of Biomedical Informatics, Biochemistry & Molecular Biophysics, and Institute for Cancer Genetics, Columbia University, 1130 Saint Nicholas Avenue, New York, NY 10032, USA
[10]Herbert Irving Comprehensive Cancer Center, Columbia University, 1130 Saint Nicholas Avenue, New York, NY 10032, USA
*Correspondence: califano@c2b2.columbia.edu
http://dx.doi.org/10.1016/j.cell.2014.09.021

SUMMARY

Identification of driver mutations in human diseases is often limited by cohort size and availability of appropriate statistical models. We propose a framework for the systematic discovery of genetic alterations that are causal determinants of disease, by prioritizing genes upstream of functional disease drivers, within regulatory networks inferred de novo from experimental data. We tested this framework by identifying the genetic determinants of the mesenchymal subtype of glioblastoma. Our analysis uncovered KLHL9 deletions as upstream activators of two previously established master regulators of the subtype, C/EBPβ and C/EBPδ. Rescue of KLHL9 expression induced proteasomal degradation of C/EBP proteins, abrogated the mesenchymal signature, and reduced tumor viability in vitro and in vivo. Deletions of KLHL9 were confirmed in > 50% of mesenchymal cases in an independent cohort, thus representing the most frequent genetic determinant of the subtype. The method generalized to study other human diseases, including breast cancer and Alzheimer's disease.

INTRODUCTION

Identification of somatic mutations and germline variants that are determinants of cancer and other complex human diseases/traits (driver mutations) is mostly performed on a statistical basis, using models of genomic evolution (Frattini et al., 2013) or mutational bias (Lawrence et al., 2013), among others, to increase the significance of individual events. Achieving appropriate statistical power, however, requires large effect sizes or large cohorts due to multiple hypothesis-testing correction (Califano et al., 2012). In addition, these approaches are not designed to provide mechanistic insight. As a result, many disease-risk determinants, such as apolipoprotein E, were discovered long before they were mechanistically elucidated (Liu et al., 2013).

Network-based analyses have recently emerged as a highly effective framework for the discovery of master regulator (MR) genes that are functional disease drivers (Aytes et al., 2014; Carro et al., 2010; Lefebvre et al., 2010; Piovan et al., 2013; Sumazin et al., 2011; Zhao et al., 2009). Here, we introduce DIGGIT (*d*river-gene *i*nference by *g*enetical-*g*enomics and *i*nformation *t*heory), an algorithm to identify genetic determinants of disease by systematically exploring regulatory/signaling networks upstream of MR genes. This collapses the number of testable hypotheses and provides regulatory clues to help elucidate associated mechanisms.

We first apply DIGGIT to identify causal genetic determinants of the mesenchymal subtype of GBM (MES-GBM), which remain poorly characterized despite extensive efforts (Brennan et al., 2013; Verhaak et al., 2010). We then demonstrate its generalizability to other diseases for which matched expression and mutational data are available.

Astrocytoma grade IV or glioblastoma (GBM) is the most common human brain malignancy and is virtually incurable, with average survival of 12–18 months post-diagnosis (Ohgaki and Kleihues, 2005). Gene-expression profile analysis revealed three subtypes associated with expression of mesenchymal, proliferative, and proneural (PN) genes, respectively (Phillips et al., 2006). Among these, mesenchymal tumors (MES-GBM) present with the worst prognosis, as confirmed by other studies (Carro et al., 2010; Sun et al., 2006; Cancer Genome Atlas Research Network, 2008). Integrative analysis of expression

and mutational data (Cancer Genome Atlas Research Network, 2008) produced a more complex stratification into PN, MES, neural, and classic subtypes, as well as into an epigenetically distinct subtype (G-CIMP) with the best prognosis (Verhaak et al., 2010). Whereas non-G-CIMP PN tumors were associated with the worst prognosis by Brennan et al. (2013), MES-GBM tumors, based on the original classification, present virtually indistinguishable prognosis and are ~7-fold more frequent (Figure S1 available online). Thus the original MES-GBM and the newer Non-G-CIMP PN signatures are both objective, equivalent markers of poor prognosis.

Among the genetic alterations reported by the TCGA Consortium (Cancer Genome Atlas Research Network, 2008), only *NF1* mutations/deletions were associated with MES-GBM tumors (~25% of samples) (Verhaak et al., 2010), although additional rare mutations and fusion events were recently reported (Danussi et al., 2013; Frattini et al., 2013). Thus, despite multiple studies, the genetic determinants of MES-GBM are still largely elusive and represent an ideal target for the new algorithm.

In Carro et al. (2010), we reported that aberrant coactivation of the transcription factors (TFs) *C/EBPβ*, *C/EBPδ*, and *STAT3* is necessary and sufficient to induce mesenchymal reprogramming in GBM, suggesting that this TF module represents an obligate pathway or *regulatory bottleneck* between driver alterations and aberrant mesenchymal program activity. We thus hypothesize that the genetic drivers of MES-GBM are either among these genes or in their upstream pathways. Use of DIGGIT to test the hypothesis elucidated two high-frequency alterations: focal amplification of *C/EBPδ* and homozygous deletion of *KLHL9*, a Cullin E3 ligase adaptor (Sumara et al., 2007).

To assess the algorithm's generalizability to other diseases and germline variants, we also applied it to breast cancer (BRCA) and Alzheimer's disease (AD). This identified driver alterations and variants missed by genome-wide association studies (GWASs) but validated by independent candidate-gene studies, as well as high-probability, yet unreported events.

RESULTS

Given a set of functional disease drivers, e.g., inferred by the MR inference algorithm (MARINa) (Aytes et al., 2014; Carro et al., 2010), DIGGIT evaluates candidate alterations in these genes and in their upstream regulators (see Figure 1A for a flowchart). This is accomplished by a five-step process (Figures 1B–1F), requiring a large set ($n \geq 200$) of gene-expression profiles (henceforth *GEPD*) to assemble and analyze regulatory networks and a large set ($n \geq 100$) of sample-matched genetic-variant profiles (henceforth *GVPD*). We first discuss application of this pipeline to identify copy-number variants (CNVs) that are causal determinants of the MES-GBM subtype. We then perform additional analyses to show that DIGGIT generalizes to the study of germline variants, as well as of other diseases, including BRCA and AD.

Step 1: MR Analysis

This step requires a context-specific regulatory network representing TF → target interactions (henceforth, *interactome*) and a gene-expression signature of interest (i.e., a p value-ranked list of differentially expressed genes) (*input*). These are analyzed by MARINa to produce a p value-ranked list of candidate MRs (*output*). Given a GEPD data set, networks can be inferred using available reverse-engineering algorithms, such as ARACNe (Basso et al., 2005). Specifically, MARINa analysis of an ARACNe-inferred GBM network, using a MES-GBM signature, identified six MR genes (MES-MRs), including *C/EBPβ*, *C/EBPδ*, *STAT3*, *BHLHB2*, *RUNX1*, and *FOSL2*, with *C/EBPβ*/*C/EBPδ* and *STAT3* as synergistic MRs (Carro et al., 2010). See Figure 1B.

Step 2: F-CNVG Analysis

Functional alterations must induce aberrant activity of their gene products (see Figure 1C). Among copy-number alterations (CNVGs), we thus select those whose ploidy is informative of gene expression as candidate functional CNVs (F-CNVGs) (Tamborero et al., 2013) (Figure S1). This is assessed based on (1) mutual information (MI) between copy number and expression or (2) differential expression in wild-type (WT) versus amplified/deleted samples (see Extended Experimental Procedures). Analyses are performed on the GEPD and sample-matched GVPD profiles (*input*), independent of subtype classification, to produce a p value-ranked list of candidate F-CNVGs (*output*).

Analysis of 229 profile-matched GBM samples in TCGA identified 1,486 candidate F-CNVGs ($p \leq 0.05$, Bonferroni corrected). The MI test proved highly sensitive, accounting for 90% of inferred F-CNVGs (Tables S1–S5) (both *KLHL9* and *C/EBPδ* were positive by MI analysis), with the t test accounting for an additional 10% of low-frequency F-CNVGs, with low MI analysis sensitivity.

Most CNVGs (94%) were thus discarded as not informative of gene expression (see Figure S1), suggesting no functional contribution. Conversely, inferred F-CNVGs included most genes previously reported as GBM drivers (14/18 > 88%) (Cancer Genome Atlas Research Network, 2008), including *EGFR*, *CDK4*, *PDGFRA*, *MDM2*, *MDM4*, *MET*, *AKT3*, *MYCN*, *PIK3CA*, *CDKN2A*, *CDKN2C*, *RB1*, *PTEN*, and *NF1* ($p = 1.2 \times 10^{-10}$) (Tables S1–S5). Analysis of remaining driver genes (*CCND2*, *CDK6*, *CDKN2B*, *PARK2*) revealed that they were missed due to either low event frequency (*CDK6* < 1.3%, *CCND2* < 2.2%, *PARK2* < 5.2%) or below-detection gene-expression levels (*CDKN2B*).

Among the MES-MRs, only *C/EBPδ* was inferred as a focally amplified F-CNVG (~22% of samples), suggesting that aberrant activity of other MES-MRs may be mediated by alterations in their upstream regulators.

Step 3: MINDy Analysis

Next, we used the MINDy algorithm to interrogate pathways upstream of MR genes (Wang et al., 2009). MINDy analyzes a large GEPD, the candidate MR list (step 1), and the F-CNVG list (step 2) (*input*) to identify F-CNVGs that are candidate post-translational modulators of MR activity (independent of subtype classification), by conditional MI analysis (Wang et al., 2009; Zhao et al., 2009; see Extended Experimental Procedures). This generates a p value-ranked list of candidate F-CNVGs in pathways upstream of MR genes (*output*). This step dramatically reduced the 1,486 F-CNVGs from step 1 to only 92 statistically significant candidate MES-MR modulators (see Table S3 and Figure 1D).