Published in final edited form as:

Hum Mutat. 2016 October ; 37(10): 1042–1050. doi:10.1002/humu.23040.

Deep genetic connection between cancer and developmental disorders

Hongjian Qi^{1,2}, Chengliang Dong^{3,4}, Wendy K. Chung⁵, Kai Wang^{3,4}, and Yufeng Shen^{2,6,7,*}

¹Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, 10027, USA

²Department of Systems Biology, Columbia University Medical Center, New York, NY, 10032, USA

³Zilkha Neurogenetic Institute, University of Southern California, Los Angeles, CA, 90089, USA

⁴Biostatistics Division, Department of Preventive Medicine, University of Southern California, Los Angeles, CA, 90089, USA

⁵Departments of Pediatrics and Medicine, Columbia University Medical Center, New York, NY, 10032, USA

⁶Department of Biomedical Informatics, Columbia University Medical Center, New York, NY, 10032, USA

⁷JP Sulzberger Columbia Genome Center, Columbia University Medical Center, New York, NY, 10032, USA

Abstract

Cancer and developmental disorders (DD) share dysregulated cellular processes such as proliferation and differentiation. There are well-known genes implicated in both in cancer and DD. In this study, we aim to quantify this genetic connection using publicly available data. We found that among DD patients, germline damaging de novo variants are more enriched in cancer driver genes than non-drivers. We estimate that cancer driver genes comprise about a third of DD risk genes. Additionally, de novo likely-gene-disrupting (LGD) variants are more enriched in tumor suppressors, and about 40% of implicated de novo damaging missense variants are located in cancer somatic mutation hotspots, indicating that many genes have a similar mode of action in cancer and DD. Our results suggest that we can view tumors as natural laboratories for assessing the deleterious effects of mutations that are applicable to germline variants and identification of causal genes and variants in DD.

*To whom correspondence may be addressed. **Corresponding authors:** Yufeng Shen, Address: 1130 St Nicholas Avenue, 812A, New York, NY 10032, Phone Number: 212-851-4662, ys2411@cumc.columbia.edu.

Author contributions

Y.S. conceived the work; Y.S. and K.W. designed the analyses; H.Q., C.D., and Y.S. conducted the analyses; H.Q., C.D., W.K.C., K.W., and Y.S. interpreted the results and wrote the manuscript.

Conflicts of Interest

K.W. is a board member and stockholder of Tute Genomics, a bioinformatics software company.

Keywords

developmental disorders; de novo mutations; somatic mutation hotspots; tumor suppressors

Introduction

De novo or rare functional variants with large effect sizes are major contributors (Neale, Kou, Liu, Ma'ayan, Samocha, Sabo, Lin, Stevens, Wang, Makarov, Polak, Yoon et al., 2012) to developmental disorders (DD), such as developmental delay, autism, intellectual disability, and epilepsy (De Rubeis, He, Goldberg, Poultney, Samocha, Cicek, Kou, Liu, Fromer, Walker, Singh, Klei et al., 2014; Deciphering Developmental Disorders, 2015; Epi, Epilepsy Phenome/Genome, Allen, Berkovic, Cossette, Delanty, Dlugos, Eichler, Epstein, Glauser, Goldstein, Han et al., 2013; Gilissen, Hehir-Kwa, Thung, van de Vorst, van Bon, Willemsen, Kwint, Janssen, Hoischen, Schenck, Leach, Klein et al., 2014; Iossifov, O'Roak, Sanders, Ronemus, Krumm, Levy, Stessman, Witherspoon, Vives, Patterson, Smith, Paepers et al., 2014). However, most of DD risk genes are still unknown: there are about 100–200 known candidate risk genes (Deciphering Developmental Disorders, 2015; Iossifov et al., 2014), yet the estimated number of risk genes that contribute to DD is about 1000 (Iossifov et al., 2014). Additionally, it is challenging to clinically interpret *de novo* or rare variants, especially missense variants, even in known risk genes.

Cancer and developmental disorders have common dysregulated cellular processes, such as proliferation, growth, and differentiation (Parikshak, Luo, Zhang, Won, Lowe, Chandran, Horvath, & Geschwind, 2013; Schubert, Shannon, & Bollag, 2007; Waite & Eng, 2003). There are well-known genes and pathways implicated in both, with recurrent somatic mutations in cancer and highly penetrant germline *de novo* variants in DD. Classic examples include *PTEN* (MIM # 601728), a negative regulator in ALK pathway implicated in autism (O'Roak, Vives, Fu, Egerton, Stanaway, Phelps, Carvill, Kumar, Lee, & Ankenman, 2012) and many types of cancer (Li, Yen, Liaw, Podsypanina, Bose, Wang, Puc, Miliareis, Rodgers, McCombie, Bigner, Giovannella et al., 1997), and *PTPN11* (MIM # 176876), a phosphatase in RAS/MAPK signaling pathway implicated in both Noonan syndrome and leukemia (Tartaglia, Niemeyer, Fragale, Song, Buechner, Jung, Hahlen, Hasle, Licht, & Gelb, 2003). Recent large-scale genomic studies of cancer (Cancer Genome Atlas Research, Weinstein, Collisson, Mills, Shaw, Ozenberger, Ellrott, Shmulevich, Sander, & Stuart, 2013; Gao, Aksoy, Dogrusoz, Dresdner, Gross, Sumer, Sun, Jacobsen, Sinha, & Larsson, 2013) and developmental disorders (de Ligt, Willemsen, van Bon, Kleefstra, Yntema, Kroes, Vulto-van Silfhout, Koolen, de Vries, Gilissen, del Rosario, Hoischen et al., 2012; De Rubeis et al., 2014; Deciphering Developmental Disorders, 2015; Epi et al., 2013; Gilissen et al., 2014; Hamdan, Srour, Capo-Chichi, Daoud, Nassif, Patry, Massicotte, Ambalavanan, Spiegelman, Diallo, Henrion, Dionne-Laporte et al., 2014; Iossifov et al., 2014; Rauch, Wieczorek, Graf, Wieland, Ende, Schwarzmayer, Albrecht, Bartholdi, Beygo, Di Donato, Dufke, Cremer et al., 2012) revealed a substantial number of genes implicated in both classes of diseases. There was reported increased burden of rare nonsynonymous variants in proto-oncogenes in autism patients (Darbro, Singh, Zimmerman, Mahajan, & Bassuk, 2016). Some of these genes share similar modes of action through cancer somatic mutations and

DD germline *de novo* variants (Ronan, Wu, & Crabtree, 2013). For example, *PTPN11* is known to harbor gain-of-function mutations that make it constitutively active in both cancer and Noonan syndrome patients; *EP300* (MIM # 602700), a tumor suppressor, has a large fraction of likely-gene-disrupting (LGD) mutations that mostly likely lead to loss-of-function in both diseases (Roelfsema, White, Ariyürek, Bartholdi, Niedrist, Papadia, Bacino, den Dunnen, van Ommen, & Breuning, 2005).

In this study, we aim to quantify the genetic connection between cancer and developmental disorders, and investigate the feasibility of utilizing cancer genomics data to help improve risk gene and variant discovery in genetic studies of developmental disorders. Driver genes are much more frequently mutated in cancer, and with ongoing international efforts in cancer precision medicine, there is an accelerated accumulation of cancer somatic mutation data. Such data will provide an unprecedented opportunity to study empirical functional consequences of mutations at virtually every base in cancer driver genes. Elucidating such a connection could lead to a better understanding of molecular mechanisms of both cancer and developmental disorders.

We compiled data sets of *de novo* variants from recently published studies on developmental disorders (De Rubeis et al., 2014; Deciphering Developmental Disorders, 2015; Epi et al., 2013; Gilissen et al., 2014; Iossifov et al., 2014; Robinson, Samocha, Kosmicki, McGrath, Neale, Perlis, & Daly, 2014), including autism, intellectual disabilities, epilepsy, and developmental delays. We also assembled a large number of candidate cancer driver genes from various sources, including Cancer Census, The Cancer Genome Atlas (TCGA) and The Candidate Cancer Gene Database (CCGD) (Abbott, Nyre, Abrahante, Ho, Vogel, & Starr, 2015; Cerami, Gao, Dogrusoz, Gross, Sumer, Aksoy, Jacobsen, Byrne, Heuer, & Larsson, 2012; Forbes, Beare, Gunasekaran, Leung, Bindal, Boutselakis, Ding, Bamford, Cole, & Ward, 2015; Gao et al., 2013). We compared the burden of *de novo* variants in candidate cancer driver genes and non-driver genes among DD cases. We then estimated the fraction of DD risk genes that are also cancer drivers. Finally, we investigated whether germline *de novo* variants and cancer somatic mutations in this set of overlapping genes have similar modes of action.

Materials and Methods

Candidate cancer driver genes

The candidate cancer driver genes list is comprised of census genes from COSMIC (Forbes et al., 2015), significantly mutated genes from TCGA studies curated by cBioPortal (Cerami et al., 2012) and candidate genes from forward genetic screens in mice in The Candidate Cancer Gene Database (CCGD) (Abbott et al., 2015) (Supp. Figure S1). For cBioPortal data, we included genes with MutSigCV q-value less than 0.1 in individual TCGA studies as significantly mutated genes. We excluded results from the Adrenocortical Carcinoma and Pancreatic Adenocarcinoma datasets because while these two cancer datasets have a moderate number of samples, many of these genes have q-values less than 0.1. For CCGD data, we only considered the genes with relative rank A (Abbott et al., 2015). We further filtered these CCGD genes based on mutation data in TCGA. Specifically, we counted the variants of various functional categories (LGD, missense, silent) reported in

TCGA (Weinstein, Collisson, Mills, Shaw, Ozenberger, Ellrott, Shmulevich, Sander, Stuart, & Network, 2013), and tested whether there is significant excess ($p < 0.05$) of missense or LGD mutations compared to silent mutations based on germline gene-specific background mutation rates (Samocha, Robinson, Sanders, Stevens, Sabo, McGrath, Kosmicki, Rehnström, Mallick, & Kirby, 2014). We note that the background somatic mutation rate is affected by various processes (Alexandrov, Nik-Zainal, Wedge, Aparicio, Behjati, Biankin, Bignell, Bolli, Borg, Borresen-Dale, Boyault, Burkhardt et al., 2013; Roberts, Lawrence, Klimczak, Grimm, Fargo, Stojanov, Kiezun, Kryukov, Carter, Saksena, Harris, Shah et al., 2013) that are different to germline mutations. However, the usage of germline background in this study is justified by the observation that there is a very strong correlation between observed ratio of missense/silent (or LGD/silent) somatic mutations and ratio of missense/silent (or LGD/silent) germline background mutation rate among non-candidate cancer driver genes (correlation coefficient = 0.46, Supp. Figure S2). All other genes not included in cancer driver genes are classified as non-cancer driver genes.

Germline *de novo* mutations of developmental disorders

We compiled germline *de novo* variants from 5542 DD cases in recent published studies, including 3962 cases with autism spectrum disorder (De Rubeis et al., 2014; Iossifov et al., 2014), 1133 cases from Deciphering Developmental Disorders study (Deciphering Developmental Disorders, 2015), 191 cases with epileptic encephalopathies (Epi et al., 2013), and 264 cases with intellectual disability (de Ligt et al., 2012; Gilissen et al., 2014; Hamdan et al., 2014; Rauch et al., 2012). We re-annotated these mutations using ANNOVAR (Wang, Li, & Hakonarson, 2010) software to have complete gene annotation as well as function annotation. The functional consequence of missense mutations is predicted *in silico* by meta-SVM. In this study, we only consider mutations in the exonic regions.

Burden test and estimation of number of causative *de novo* mutations

Burden test between case and control—Within gene sets, we counted the number of mutations inside the gene set of different functional categories (LGD, missense, silent) for both cases and controls. We assumed that *de novo* variants are sequences of individual Bernoulli trials and we used the portion of case trios as the success probability to calculate the two-side binomial distribution p-value as well as fold enrichment.

Burden comparison between cancer and non-cancer drivers—We counted the number of *de novo* variants in candidate cancer driver genes and non-cancer drivers of different functional categories (LGD, missense, silent) for both cases and controls. We used two-side Fisher's exact test to test the null hypothesis that the case/control burden of various categories is the same among cancer driver genes and non-cancer driver genes.

Estimation of number of causative *de novo* mutations and class vulnerability—(Iossifov et al., 2014) in gene sets: In a group of genes (e.g. cancer drivers), there are L_1 LGD (or D-mis) mutations from n_1 cases and L_2 LGD (or D-mis) mutations from n_2 controls, we estimate the number of *causative variants* C by:

$$C=L_1-\frac{L_2*n_1}{n_2}$$

and *class vulnerability* V by:

$$V=\frac{C}{L_1}$$

Using the L_1 and L_2 as the Poisson distribution rate to simulate 10,000 trials, we can calculate the 95% confident intervals of *causative variants* and *class vulnerability*.

Estimation of percentage of causative mutations in cancer driver genes—We first counted the number of *de novo* LGD (or D-mis) variants in all genes and in candidate cancer driver genes for both cases and controls, then we used the variant counts as the Poisson distribution rate to simulate 10,000 trials. Dividing the number of simulated causative variants in cancer driver genes by the simulated mutations in all genes, we obtain the expectation as well as a 95% confident interval.

Infer candidate risk genes of DD

TADA (transmission and *de novo* association) (De Rubeis et al., 2014; He, Sanders, Liu, De Rubeis, Lim, Sutcliffe, Schellenberg, Gibbs, Daly, & Buxbaum, 2013) is a Bayesian method for identification of risk genes using rare or *de novo* variants. We tallied the occurrence of *de novo* variants in two categories: LGD and D-mis. We used gene-specific mutation rate (Samocha et al., 2014) as the parameter for the Poisson distribution and calculated its corresponding FDR using other default parameters.

DD risk genes that overlap with candidate cancer driver genes—We defined DD candidate risk genes using FDR calculated by TADA. With each FDR threshold, we obtained the number of candidate DD risk genes (N) and the number (N_c) of such genes that are also candidate cancer drivers. We estimated the number (F) of false positive DD risk genes by FDR definition: $F=N*FDR$. To estimate the fraction (f) of true DD risk genes that overlap with candidate cancer driver genes, we assumed false positive DD risk genes overlap with candidate cancer driver genes just by chance, which is determined by background germline *de novo* mutation rate. In most TADA FDR bins ($FDR < 0.5$), the false positive risk genes should have at least one damaging *de novo* variant (LGD or D-mis). By calculating the sum of germline damaging mutation rate in cancer driver genes divided by all genes, we determined that the overlap rate by chance is $r=10\%$. Finally, for each TADA FDR bin, the fraction of true DD risk genes that are also candidate cancer driver genes was estimated by:

$$f=\frac{N_c-F*r}{N-F}$$

A Hidden Markov Model to infer cancer somatic missense mutation hotspots

We implemented a hidden Markov Model (HMM) to predict somatic missense mutation hotspots in each candidate cancer driver gene. We assume that the background somatic mutation rate is uniform across a given gene. For each transcript in the given gene, we inspected the somatic missense mutations from COSMIC. We counted all missense mutations at each given amino acid site, regardless of actually amino acid changes, to identify mutation hotspots. We defined hotspots in two ways: 1) highly-recurrent mutation sites and 2) sites with non-background states prediction by HMM. Recurrent mutated positions were defined as having more than 3.5 median-absolute-deviation number of mutations. After excluding recurrent sites, we took a sliding window of size 8 and summed the number of mutations for each position to reduce the fluctuation of mutations in a neighborhood region. We used the smoothed position-specific mutation counts as the input to a Hidden Markov Model with Poisson emission probability and three hidden states, including: (a) the “background” state, (b) possible mutation hotspot state, and (c) probable mutation hotspot state. We used germline mutation background to estimate the fraction of missense mutations that are drivers in each gene. This is based on the observation that, among non-cancer driver genes, the ratio of reported missense/silent somatic mutations is close to gene-specific background mutation rate estimated by Samocha et al 2014 (regression slope = 0.97 and intercept close to zero, Supp. Figure S2.). We simulated the missense mutation counts (S_1) and silent mutation counts (S_2) in each gene using the corresponding recorded COSMIC data (C_1 , C_2) as the Poisson distribution rate. With the missense/silent ratio from germline *de novo* background (R_1), we estimated the mean and 95% confidence interval of the fraction of missense mutations (f) that are drivers in each gene by:

$$f = \frac{S_1 - S_2 * R_1}{S_1}$$

We used the upper bound of 95% CI as the maximum allowed (M) percentage of missense drivers from HMM. To obtain reasonable initial values for HMM parameters, we then calculated the expected number of driver missense mutations per position (T) by:

$$T = \frac{C_1 * (1 - f)}{L}$$

where L is the total transcript length. We set the lambda (mean of a Poisson distribution) for the background state to be at least T . To restrict the number of transitions between background and hotspot states, we took the average of the diagonal of the transition matrix of the Baum-Welsh result with 0.99 if the corresponding transition matrix elements were smaller than 0.99 in each iteration. After convergence, we used the Viterbi algorithm to find the most probable state path and forward-background algorithm to calculate posterior marginal probabilities of hidden states for each position. To identify the hotspots, we took positions with the non-background states as the hotspots, with exception that if the fraction of somatic missense mutations in those hotspots exceeded M , we ranked those positions by

their marginal probability of being background states (increasingly), and included such positions until the fraction of missense mutations in hotspots reached M .

Results

Burden of germline *de novo* variants in developmental disorder patients among candidate cancer driver genes

To investigate the contribution of cancer driver genes to developmental disorders (DD), we compiled a large dataset of 6294 germline *de novo* coding variants from 5542 DD cases drawn from recent published studies, including 3953 cases with autism spectrum disorder (De Rubeis et al., 2014; Iossifov et al., 2014), 1133 cases with various developmental disorders from Deciphering Developmental Disorders study (Deciphering Developmental Disorders, 2015), 192 cases with epileptic encephalopathies (Epi et al., 2013), and 264 cases with intellectual disability (de Ligt et al., 2012; Gilissen et al., 2014; Hamdan et al., 2014; Rauch et al., 2012) (Supp. Table S1, S2). Aggregating various DD with shared but distinct genetic risk architectures can yield additional findings in risk genes. (Deciphering Developmental Disorders, 2015; Homsy, Zaidi, Shen, Ware, Samocha, Karczewski, DePalma, McKean, Wakimoto, & Gorham, 2015; Retterer, Juusola, Cho, Vitazka, Millan, Gibellini, Vertino-Bell, Smaoui, Neidich, & Monaghan, 2015) We re-annotated these variants using ANNOVAR software (Wang et al., 2010), and predicted the functional consequences of missense variants *in silico* using meta-SVM (Dong, Wei, Jian, Gibbs, Boerwinkle, Wang, & Liu, 2015). The following analyses are focused on likely-gene-disrupting (LGD, which includes stopgain, stoploss, frameshifting and splicing variants) or predicted-damaging missense (D-mis, predicted by meta-SVM) variants. We used 1911 parents-unaffected sibling trios from the Simons Simplex Collection (SSC) as controls (Iossifov et al., 2014). The overall rate of silent *de novo* variants is similar between cases and controls (0.25 per subject; Supp. Table S3). To include a broad set of cancer driver genes (Cheng, Zhao, & Zhao, 2015), we obtained 568 cancer census genes from COSMIC (Forbes et al., 2015), 773 genes with MutSigCV (Lawrence, Stojanov, Polak, Kryukov, Cibulskis, Sivachenko, Carter, Stewart, Mermel, Roberts, Kiezun, Hammerman et al., 2013) *q*-value less than 0.1 from individual The Cancer Genome Atlas (TCGA) studies curated by cBioPortal (cBioPortal, 2015; Gao et al., 2013), and 325 candidate driver genes from forward genetic screens in mice by The Candidate Cancer Gene Database (CCGD) (Abbott et al., 2015) (see methods section). In total, we compiled a list of 1481 candidate cancer driver genes (Supp. Table S4, Supp. Figure S1), all other genes not classified as cancer driver genes were considered as non-cancer driver genes.

Among all candidate cancer driver genes, there is significant enrichment of LGD or D-mis germline *de novo* variants in DD cases compared to controls (Supp. Table S5(a)), and such enrichment is significant in both autism and other types of NDD (Supp. Table S5(b)). Moreover, among DD cases, candidate cancer driver genes show significantly higher enrichment of germline *de novo* variants than non-drivers (Table 1, odds ratio = 2.0, *p*-value = 4.5e-6). Such enrichment cannot simply be explained by known constrained genes in cancer driver genes, as we observed that among constrained genes (Lek, Karczewski, Minikel, Samocha, Banks, Fennell, O'Donnell-Luria, Ware, Hill, & Cummings, 2015) there

is still a significantly greater burden in drivers than in non-drivers (odds ratio = 2.1, Supp. Table S5(c)). Based on fold enrichment of *de novo* variants in cases compared to controls, we estimate that there are about 391 causative LGD variants in total, of which 163 are in candidate cancer driver genes (42%), and that there are about 327 causative D-mis variants in total, of which 114 are located in cancer driver genes (35%). Therefore, about 38% (95% confidence interval (CI): [29%, 51%]) of all potentially causative damaging (LGD or D-mis) *de novo* variants observed in these DD cases are located in candidate cancer drivers.

Among the candidate cancer drivers that harbor damaging *de novo* variants in DD cases, several pathways are enriched (Supp. Table S6), including transcriptional regulation (e.g. lysine degradation), core developmental pathways (e.g. Wnt and Hippo signaling), pathways related to cell junctions and adhesion, and ubiquitin mediated proteolysis.

Cancer driver genes comprise about a third of DD risk genes

To identify a broad set of candidate risk genes of developmental disorders, we applied TADA (transmission and *de novo* association) (De Rubeis et al., 2014; He et al., 2013), a probabilistic method for identifying risk genes based on *de novo* or rare variants. We used only *de novo* LGD and D-mis mutations in this analysis and gene-specific background mutation rate (Samocha et al., 2014; Ware, 2015). We ranked all genes by false discovery rate (FDR), defined DD candidate risk genes using increasing thresholds of 10% bins, and then calculated the percentage of true DD risk genes that overlap with cancer driver genes (Table 2). The estimated overlap percentage is 45% at FDR 0.1; the ratio decreases at larger FDR values but is still significantly greater than what is expected by chance at FDR of 0.5. Overall, we estimate that cancer drivers comprise more than a third of risk genes contributing to developmental diseases.

Germline *de novo* variants disrupt DD risk genes through similar modes of action as somatic mutations in cancer drivers

Cancer driver genes are generally categorized as tumor suppressors or oncogenes, with the exception of genes that play either role in different cancer types (Lobry, Oh, & Aifantis, 2011). The molecular consequence of a driver somatic mutation is usually loss-of-function in a tumor suppressor gene and gain-of-function in an oncogene. There are a number of known DD risk genes disrupted by germline variants via similar modes of action as cancer driver genes disrupted by somatic mutations. For example, gain-of-function germline variants in *SOS1* (MIM # 182530) and *PTPN11* genes are implicated in Noonan syndrome (Tartaglia, Mehler, Goldberg, Zampino, Brunner, Kremer, van der Burgt, Crosby, Ion, Jeffery, Kalidas, Patton et al., 2001); Both genes are also oncogenes with gain-of-function somatic mutations in leukemia (Tartaglia et al., 2003). To quantify the similarity of modes of action between cancer and DD in individual genes and pathways, we investigated the patterns of cancer somatic mutations and DD germline *de novo* variants. We made two assumptions: (a) loss-of-function mutations include both truncating mutations (LGD, including stopgain, stoploss, splicing, and frameshifting), and a subset of missense mutations. Tumor suppressors tend to harbor both types of mutations, generally with a large fraction of LGD mutations (Davoli, Xu, Mengwasser, Sack, Yoon, Park, & Elledge, 2013);

(b) gain-of-function mutations are mostly composed of missense mutations. We note that genes with dominant negative mutations are often exceptions.

We reasoned that tumor suppressors are likely haploinsufficient (Davoli et al., 2013) as DD risk genes. To test that, we identified likely tumor suppressor genes and likely non-suppressor genes based on the fraction of LGD mutations among all somatic SNVs and indels in a given gene, across all cancers. Specifically, we grouped the candidate cancer driver genes into 4 bins using data from COSMIC, with fractions of LGD mutations at 0–5%, 5–10%, 10–15%, and 15% or larger, respectively. Among these bins, the genes in the 0–5% bin are likely non-suppressors, and the ones in the 15% or larger bin are likely tumor suppressor genes. This tier classification is consistent with an independent study of predicted tumor suppressor genes (Davoli et al., 2013) using TCGA data. More than half of the likely tumor suppressor genes overlap with the predicted confident tumor suppressor genes; there is almost no overlap of likely non-suppressor with predicted confident tumor suppressor genes (Davoli et al., 2013) (Supp. Figure S3). Compared to the number of germline LGD *de novo* variants expected from background mutation rate (Samocha et al., 2014; Ware, 2015) in DD cases, we observed a 2.4 enrichment in likely non-suppressor genes (Supp. Table S7), which, as expected, is below the overall enrichment in cancer driver genes (3.3x, Supp. Table S5). On the contrary, we observed more than 10 fold enrichment of LGD variants in likely-tumor suppressor genes than expected (Figure 1; Supp. Table S7), representing a 4.2x greater enrichment than in likely non-suppressors. This indicates that tumor suppressor genes implicated in DD patients through germline *de novo* variants often confer disease risk via loss of function.

Functional missense mutations, whether gain or loss of function, disrupt cellular processes in very specific ways. For example, these mutations can cause (gain or loss of) enzymatic activity or (loss of) regulation of protein stability/activity, or affect interaction with other proteins. Therefore, functional missense mutations tend to form clusters in specific regions. We denote these clusters as cancer mutation hotspots. We found that for amino acid positions where there were at least 3 reported somatic missense mutations in COSMIC, there are 34 *de novo* missense variants in cases and just 1 in controls (fold enrichment = 12, p-value = 6.9e-4, Supp. Table S8). There was a consistent trend among *de novo* D-mis variants at positions with 1 or 2 reported somatic missense mutations (fold enrichment = 3.1, p-value = 0.028, Supp. Table S8). Several methods have been developed to find mutation hotspots for the purpose of finding cancer driver genes with a high accuracy (Chang, Asthana, Gao, Lee, Chapman, Kandath, Gao, Socci, Solit, & Olshen, 2015; Yang, Petsalaki, Rolland, Hill, Vidal, & Roth, 2015). Among the reported cancer mutation hotspots, we observed a similar trend of enriched *de novo* mutations in DD cases (Supp. Table S9). To reach optimal power for this study with a balance of accuracy and sensitivity, we implemented a hidden Markov Model (HMM) to predict these hotspots (Methods section) in genes that are already implicated as candidate drivers. We collected all somatic missense mutations from COSMIC for each gene and applied our HMM-based methods to detect missense mutation hotspots in all candidate cancer driver genes (Supp. Table S10). Comparing DD cases with controls, we observed a 16x fold enrichment (p-value = 1.8e-5) of germline *de novo* D-mis variants in cancer mutation hotspots (Table 3, Supp. Table S11, S12), which indicates that almost all such mutations contribute to DD and corresponds to a class vulnerability value of 90%

(Iossifov et al., 2014), much greater than D-mis variants (about 25%) in non-cancer drivers (Figure 2). Based on fold enrichment, the estimated number of DD-causative *de novo* missense variants among all candidate cancer drivers is about 135, and the estimated number of such variants in cancer somatic mutation hotspots is 67. This suggests that a large portion (about 50%) of causative *de novo* missense variants in DD cases among cancer driver genes have similar modes of action as cancer somatic mutations.

Figure 3 shows a few representative genes. *EP300*, a known tumor suppressor, has one D-mis *de novo* variant at a cancer mutation hotspot in one DD patient (Figure 3a), consistent with its implicated role through loss of function with other 5 LGD *de novo* variants in DD data sets. *ARID2* (MIM # 609539) is another tumor suppressor and part of SWI/SNF chromatin remodeling complex. One autism patient had a germline *de novo* missense variant in *ARID2* at a somatic mutation hotspot (Figure 3b). *DDX3X* (MIM # 300160), a tumor suppressor implicated in intellectual disability (Deciphering Developmental Disorders, 2015), has three missense *de novo* variants in our compiled DD data sets and all located in cancer hotspots (Figure 3c). *MAP2K1* (MIM # 176872), a proto-oncogene, has a missense *de novo* variant located in a cancer hotspot (Figure 3d) in an autism case, suggesting that the variant plays a similar role as gain-of-function mutations implicated in syndromes (Rodriguez-Viciana, Tetsu, Tidyman, Estep, Conger, Santa Cruz, McCormick, & Rauen, 2006) with ASD features. *PPP2R1A* (MIM # 605983), a recently discovered DD risk gene (Deciphering Developmental Disorders, 2015), harbors three missense *de novo* variants in two cancer hotspots (Figure 3e). *PPP2R1A* is likely an oncogene in ovarian clear cell carcinoma (Jones, Wang, Shih Ie, Mao, Nakayama, Roden, Glas, Slamon, Diaz, Vogelstein, Kinzler, Velculescu et al., 2010), consistent with its gain of function roles in both cancer and DD. *CNOT3* (MIM # 604910), a tumor suppressor gene (De Keersmaecker, Atak, Li, Vicente, Patchett, Girardi, Gianfelici, Geerdens, Clappier, & Porcu, 2013), has two *de novo* LGD variants and two *de novo* D-mis variants, one of which is located in a mutation hotspot (Figure 3f), indicating it is a potential DD risk gene. Not all driver genes have similar mode of action. *CTNNB1* (MIM # 116806), a central player in Wnt signaling, is a proto-oncogene in various cancers (Morin, 1999). It has a very small fraction (0.5%) of LGD somatic mutations in COSMIC, and most missense somatic mutations disrupt the phosphorylation sites at the N-terminal end that are required for phosphorylation-dependent degradation. In contrast to somatic mutations in cancer, it is usually haploinsufficient and harbors LGD variants in patients with neurodevelopmental syndromes (Tucci, Kleefstra, Hardy, Heise, Maggi, Willemsen, Hilton, Esapa, Simon, & Buenavista, 2014). In the DD data sets we compiled there are 7 LGD *de novo* variants, consistent with a haploinsufficiency mechanism. In addition, there is a missense variant in an autism case. This missense variant is not located in any somatic mutation hotspot (Figure 3g), and is therefore unlikely to cause gain-of-function in *CTNNB1*. This is consistent with the notion that, this variant is either implicated in autism via loss of function similar to other LGD variants, or not associated with the disease. *SMARCA4* (MIM # 603254), a tumor suppressor gene (Medina, Romero, Kohno, Montuenga, Pio, Yokota, & Sanchez-Cespedes, 2008; Vandeweyer, Helsmoortel, Van Dijck, Vulto-van Silfhout, Coe, Bernier, Gerds, Rooms, van den Ende, & Bakshi, 2014), harbors three deleterious missense *de novo* mutations in the NDD cases, none of which is located in cancer mutation hotspots (Figure 3h). This is consistent with previous

report that *SMARCA4* may have gain of function or dominant negative mutations in developmental disorders (Tsurusaki, Okamoto, Ohashi, Kosho, Imai, Hibi-Ko, Kaname, Naritomi, Kawame, Wakui, Fukushima, Homma et al., 2012).

Discussion

Recent large-scale exome sequencing studies of developmental disorders (DD) uncovered many candidate risk genes and pathways through deleterious germline *de novo* mutations. Many of these genes and pathways have been previously implicated in cancer through somatic mutations. Such genetic connection is reasonable because both classes of diseases involve disruption of similar fundamental cellular processes such as growth, proliferation, and differentiation. In this study we hypothesize that quantifying such connection between DD and cancer would lead to better understanding of how genes are disrupted through mutations, and ultimately allow us to leverage the vast amount of cancer mutation data to improve genetic discovery in DD studies. Based on data from recently published large-scale DD studies and cancer genomics resources, we found that in DD patients there is a significantly greater burden of functional *de novo* mutations in candidate cancer driver genes than in non-cancer driver genes. And such enrichment trend holds in both candidate tumor suppressors and oncogenes (Supp. Table S13 and Supp. Note). Specifically, about 38% of all potentially causative damaging *de novo* mutations observed in these DD patients are located in cancer drivers, and about 27–45% of DD risk genes are likely cancer driver genes. This indicates that we can prioritize known cancer driver genes to find candidate risk genes in DD studies.

Additionally, we investigated whether driver somatic mutations in cancer and causative germline *de novo* variants in DD have similar modes of action. We found that likely tumor suppressor genes, i.e., the ones with larger fraction of LGD mutations (15%) reported in COSMIC, have a significantly higher burden of germline *de novo* LGD variants than likely non-suppressors (somatic LGD fraction < 5%) in DD patients, indicating that tumor suppressor genes often exert DD risk through loss of function germline *de novo* variants that disrupt molecular pathways in DD similar to the ones in cancer. Several well-known oncogenes have gain-of-function germline *de novo* missense variants that cause DD. However, in general it remains a challenge to infer whether a missense mutation causes gain or loss of function of the gene. We therefore asked whether missense mutations in cancer (somatic) driver genes and DD germline risk genes have similar modes of action. We hypothesized that in both tumor suppressors and oncogenes, functional somatic missense mutations in driver genes occur in “hotspots” in a driver gene. We indeed found a stronger enrichment of damaging *de novo* missense variants located in these hotspots in DD patients than in controls. Specifically, we estimated that about 48% of causative *de novo* missense variants observed in DD among cancer driver genes are located in hotspots, indicating that missense mutations also often have similar mode of action in cancer and DD. This ratio is likely an under-estimate, since the power of detecting mutation hotspots is limited in many cancer driver genes due to relatively small number of mutations. We observed this enrichment of hotspot variants in both tumor suppressors and oncogenes. Strikingly, the case-control comparison indicates that when located in cancer mutation hotspots, most of the germline *de novo* missense mutations in DD patients are implicated with the disease. This

suggests that in addition to using cancer driver genes to prioritize candidate risk gene in DD studies, we can leverage cancer somatic mutation data to improve functional assessment of germline rare or *de novo* variants in these genes observed in DD patients, potentially improving both risk gene discovery in genetic studies and genetic diagnosis in clinical testing.

Based on the evidence of LGD variants in tumor suppressors and D-mis variants located in cancer somatic mutation hotspots, we identified two new candidate risk genes for DD. The first is *ARID2*, which harbors a *de novo* D-mis variant in an autism patient. The variant is located at the second most recurrently mutated position reported in COSMIC. A potential role of *ARID2* in autism is consistent with its recently implicated role in causing intellectual disability with *de novo* LGD variants (Shang, Cho, Retterer, Folk, Humberson, Rohena, Sidhu, Saliganan, Iglesias, Vitazka, Juusola, O'Donnell-Luria et al., 2015). The second gene is *CNOT3*, which harbors two *de novo* LGD variants and two *de novo* D-mis variants in four different patients, including one with autism and three with undiagnosed developmental disorders (Deciphering Developmental Disorders, 2015). *CNOT3* is a tumor suppressor (De Keersmaecker et al., 2013) with a very large fraction (~24% (Forbes et al., 2015)) of LGD mutations among all reported somatic mutations, indicating that its suppressor role is through haploinsufficiency (Davoli et al., 2013). One of the *de novo* D-mis variants is located at the most recurrently mutated site reported in COSMIC. *CNOT3* is a component of CCR4-NOT complex, which is one of the major cellular mRNA deadenylases (Albert, Lemaire, van Berkum, Gentz, Collart, & Timmers, 2000) and has a broad role in post-transcriptional regulation of gene expression (Chen, Chiang, & Denis, 2002). Post-transcriptional regulation of gene expression has been implicated as a major pathway with neurodevelopment disorders (De Rubeis et al., 2014). This supports *CNOT3* as a candidate risk gene of developmental disorders. Future genetic and functional studies are required to confirm and validate these two candidate risk genes.

In summary, our results suggest that we can view tumors as natural laboratories for assessing the deleterious effects of mutations that are applicable to germline variants, which will enable us to improve identification of causal genes and variants in DD. Our study is still limited by inadequate number of sequenced cancer genomes in a few ways. First, we have limited power to detect mutation hotspots in a substantial portion of cancer driver genes due to a relatively small number of mutations, especially among the genes that are mutated in a small fraction of cancer patients or cancer types. This lack of power leads to lower sensitivity and specificity. Recent works on clustering of somatic mutations in 3D (Kamburov, Lawrence, Polak, Leshchiner, Lage, Golub, Lander, & Getz, 2015) or pooled homologous domains (Miller, Rialdi, Ho, Tilove, Martinez-Gil, Moshkina, Peralta, Noel, Melegari, & Maestre, 2015) present promising directions to improve the power. Second, many genes have a diverse set of functions, and clinically distinct types of cancer (Hanahan & Weinberg, 2011; Muller & Vousden, 2013) or diseases often involve disruption of different functions of the same gene. Categorizing these disruptions as gain or loss of function is overly simplification. While our approach of detecting somatic mutation hotspots does not rely on such simplified assumption, the complexity does lead to decreased power in detection of somatic mutation hotspots, and increased difficulty in utilization and interpretation of the somatic mutation hotspots in developmental disorders. Finally, we do

not have a complete catalog of cancer driver genes, and our list of candidate cancer driver genes may contain a non-negligible number of false positives. Ongoing international efforts in cancer precision medicine are generating much larger cancer mutation data sets. With prudent data sharing practices, this will improve cancer driver genes and mutation hotspots detection in the future, and make cancer data more valuable to genetic studies and diagnosis of developmental disorders.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Jonathan Packer, Qiang Huang, Alexander Hsieh, and John J. Long for helpful discussions. This work was supported by NIH grant R01HG008157 (H.Q. and Y.S.) and U01HG008680 (Y.S.).

References

- Abbott KL, Nyre ET, Abrahante J, Ho Y-Y, Vogel RI, Starr TK. The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic acids research*. 2015; 43(1D):D844–D848. [PubMed: 25190456]
- Albert TK, Lemaire M, van Berkum NL, Gentz R, Collart MA, Timmers HT. Isolation and characterization of human orthologs of yeast CCR4-NOT complex subunits. *Nucleic Acids Res*. 2000; 28(3):809–817. [PubMed: 10637334]
- Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, et al. Signatures of mutational processes in human cancer. *Nature*. 2013; 500(7463):415–421. [PubMed: 23945592]
- Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. Cancer Genome Atlas Research, N. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013; 45(10):1113–1120. [PubMed: 24071849]
- cBioPortal. [Retrieved 12 Apr, 2015] Data sets of published studies. 2015. from http://www.cbioportal.org/data_sets.jsp
- Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*. 2012; 2(5):401–404. [PubMed: 22588877]
- Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandath C, Gao J, Socci ND, Solit DB, Olshen AB. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature biotechnology*. 2015
- Chen J, Chiang YC, Denis CL. CCR4, a 3'-5' poly(A) RNA and ssDNA exonuclease, is the catalytic component of the cytoplasmic deadenylase. *EMBO J*. 2002; 21(6):1414–1426. [PubMed: 11889047]
- Cheng F, Zhao J, Zhao Z. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Briefings in Bioinformatics*. 2015 bbv068.
- Darbro BW, Singh R, Zimmerman MB, Mahajan VB, Bassuk AG. Autism Linked to Increased Oncogene Mutations but Decreased Cancer Rate. *PLoS One*. 2016; 11(3):e0149041. [PubMed: 26934580]
- Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, Elledge SJ. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*. 2013; 155(4):948–962. [PubMed: 24183448]
- De Keersmaecker K, Atak ZK, Li N, Vicente C, Patchett S, Girardi T, Gianfelici V, Geerdens E, Clappier E, Porcu M. Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. *Nat Genet*. 2013; 45(2):186–190. [PubMed: 23263491]

- de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C, del Rosario M, Hoischen A, et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med*. 2012; 367(20):1921–1929. [PubMed: 23033978]
- De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, Kou Y, Liu L, Fromer M, Walker S, Singh T, Klei L, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*. 2014; 515(7526):209–215. [PubMed: 25363760]
- Deciphering Developmental Disorders, S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature*. 2015; 519(7542):223–228. [PubMed: 25533962]
- Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet*. 2015; 24(8):2125–2137. [PubMed: 25552646]
- Epi KC, Epilepsy Phenome/Genome P, Allen AS, Berkovic SF, Cossette P, Delanty N, Dlugos D, Eichler EE, Epstein MP, Glauser T, Goldstein DB, Han Y, et al. De novo mutations in epileptic encephalopathies. *Nature*. 2013; 501(7466):217–221. [PubMed: 23934111]
- Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research*. 2015; 43(D1):D805–D811. [PubMed: 25355519]
- Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling*. 2013; 6(269):p11–p11. [PubMed: 23550210]
- Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, Willemsen MH, Kwint M, Janssen IM, Hoischen A, Schenck A, Leach R, Klein R, et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature*. 2014; 511(7509):344–347. [PubMed: 24896178]
- Hamdan FF, Srour M, Capo-Chichi JM, Daoud H, Nassif C, Patry L, Massicotte C, Ambalavanan A, Spiegelman D, Diallo O, Henrion E, Dionne-Laporte A, et al. De novo mutations in moderate or severe intellectual disability. *PLoS Genet*. 2014; 10(10):e1004772. [PubMed: 25356899]
- Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011; 144(5):646–674. [PubMed: 21376230]
- He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, Schellenberg GD, Gibbs RA, Daly MJ, Buxbaum JD. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet*. 2013; 9(8):e1003671. [PubMed: 23966865]
- Homsy J, Zaidi S, Shen Y, Ware JS, Samocha KE, Karczewski KJ, DePalma SR, McKean D, Wakimoto H, Gorham J. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science*. 2015; 350(6265):1262–1266. [PubMed: 26785492]
- Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, Stessman HA, Witherspoon KT, Vives L, Patterson KE, Smith JD, Paepers B, et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature*. 2014; 515(7526):216–221. [PubMed: 25363768]
- Jones S, Wang TL, Shih Ie M, Mao TL, Nakayama K, Roden R, Glas R, Slamon D, Diaz LA Jr, Vogelstein B, Kinzler KW, Velculescu VE, et al. Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science*. 2010; 330(6001):228–231. [PubMed: 20826764]
- Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, Lander ES, Getz G. Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proceedings of the National Academy of Sciences*. 2015; 112(40):E5486–E5495.
- Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013; 499(7457):214–218. [PubMed: 23770567]
- Lek M, Karczewski K, Minikel E, Samocha K, Banks E, Fennell T, O'Donnell-Luria A, Ware J, Hill A, Cummings B. Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*. 2015 030338.
- Li J, Yen C, Liaw D, Podsypanina K, Bose S, Wang SI, Puc J, Miliaresis C, Rodgers L, McCombie R, Bigner SH, Giovanella BC, et al. PTEN, a putative protein tyrosine phosphatase gene mutated in

- human brain, breast, and prostate cancer. *Science*. 1997; 275(5308):1943–1947. [PubMed: 9072974]
- Lobry C, Oh P, Aifantis I. Oncogenic and tumor suppressor functions of Notch in cancer: it's NOTCH what you think. *J Exp Med*. 2011; 208(10):1931–1935. [PubMed: 21948802]
- Medina PP, Romero OA, Kohno T, Montuenga LM, Pio R, Yokota J, Sanchez-Cespedes M. Frequent BRG1/SMARCA4-inactivating mutations in human lung cancer cell lines. *Human mutation*. 2008; 29(5):617–622. [PubMed: 18386774]
- Miller MS, Rialdi A, Ho JSY, Tilove M, Martinez-Gil L, Moshkina NP, Peralta Z, Noel J, Melegari C, Maestre AM. Senataxin suppresses the antiviral transcriptional response and controls viral biogenesis. *Nature immunology*. 2015; 16(5):485–494. [PubMed: 25822250]
- Morin PJ. β -catenin signaling and cancer. *Bioessays*. 1999; 21(12):1021–1030. [PubMed: 10580987]
- Muller PA, Vousden KH. p53 mutations in cancer. *Nature cell biology*. 2013; 15(1):2–8. [PubMed: 23263379]
- Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, Polak P, Yoon S, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*. 2012; 485(7397):242–245. [PubMed: 22495311]
- O'Roak BJ, Vives L, Fu W, Egerton JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C, Ankenman K. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science*. 2012; 338(6114):1619–1622. [PubMed: 23160955]
- Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, Horvath S, Geschwind DH. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*. 2013; 155(5):1008–1021. [PubMed: 24267887]
- Rauch A, Wieczorek D, Graf E, Wieland T, Ende S, Schwarzmayr T, Albrecht B, Bartholdi D, Beygo J, Di Donato N, Dufke A, Cremer K, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet*. 2012; 380(9854):1674–1682. [PubMed: 23020937]
- Retterer K, Juusola J, Cho MT, Vitazka P, Millan F, Gibellini F, Vertino-Bell A, Smaoui N, Neidich J, Monaghan KG. Clinical application of whole-exome sequencing across clinical indications. *Genetics in Medicine*. 2015
- Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G, Harris S, Shah RR, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet*. 2013; 45(9):970–976. [PubMed: 23852170]
- Robinson EB, Samocha KE, Kosmicki JA, McGrath L, Neale BM, Perlis RH, Daly MJ. Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. *Proc Natl Acad Sci U S A*. 2014
- Rodriguez-Viciano P, Tetsu O, Tidyman WE, Estep AL, Conger BA, Santa Cruz M, McCormick F, Rauhen KA. Germline mutations in genes within the MAPK pathway cause cardio-facio-cutaneous syndrome. *Science*. 2006; 311(5765):1287–1290. [PubMed: 16439621]
- Roelfsema JH, White SJ, Ariyürek Y, Bartholdi D, Niedrist D, Papadia F, Bacino CA, den Dunnen JT, van Ommen G-JB, Breuning MH. Genetic heterogeneity in Rubinstein-Taybi syndrome: mutations in both the CBP and EP300 genes cause disease. *The American Journal of Human Genetics*. 2005; 76(4):572–580. [PubMed: 15706485]
- Ronan JL, Wu W, Crabtree GR. From neural development to cognition: unexpected roles for chromatin. *Nat Rev Genet*. 2013; 14(5):347–359. [PubMed: 23568486]
- Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnström K, Mallick S, Kirby A. A framework for the interpretation of de novo mutation in human disease. *Nat Genet*. 2014; 46(9):944–950. [PubMed: 25086666]
- Schubert S, Shannon K, Bollag G. Hyperactive Ras in developmental disorders and cancer. *Nature Reviews Cancer*. 2007; 7(4):295–308. [PubMed: 17384584]
- Shang L, Cho MT, Retterer K, Folk L, Humberson J, Rohena L, Sidhu A, Saliganan S, Iglesias A, Vitazka P, Juusola J, O'Donnell-Luria AH, et al. Mutations in ARID2 are associated with intellectual disabilities. *Neurogenetics*. 2015
- Tartaglia M, Mehler EL, Goldberg R, Zampino G, Brunner HG, Kremer H, van der Burgt I, Crosby AH, Ion A, Jeffery S, Kalidas K, Patton MA, et al. Mutations in PTPN11, encoding the protein

- tyrosine phosphatase SHP-2, cause Noonan syndrome. *Nat Genet.* 2001; 29(4):465–468. [PubMed: 11704759]
- Tartaglia M, Niemeyer CM, Fragale A, Song X, Buechner J, Jung A, Hahlen K, Hasle H, Licht JD, Gelb BD. Somatic mutations in PTPN11 in juvenile myelomonocytic leukemia, myelodysplastic syndromes and acute myeloid leukemia. *Nat Genet.* 2003; 34(2):148–150. [PubMed: 12717436]
- Tsurusaki Y, Okamoto N, Ohashi H, Kosho T, Imai Y, Hibi-Ko Y, Kaname T, Naritomi K, Kawame H, Wakui K, Fukushima Y, Homma T, et al. Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. *Nat Genet.* 2012; 44(4):376–378. [PubMed: 22426308]
- Tucci V, Kleefstra T, Hardy A, Heise I, Maggi S, Willemsen MH, Hilton H, Esapa C, Simon M, Buenavista M-T. Dominant β -catenin mutations cause intellectual disability with recognizable syndromic features. *The Journal of clinical investigation.* 2014; 124(4):1468. [PubMed: 24614104]
- Vandeweyer, G.; Helsmoortel, C.; Van Dijck, A.; Vulto-van Silfhout, AT.; Coe, BP.; Bernier, R.; Gerds, J.; Rooms, L.; van den Ende, J.; Bakshi, M. The transcriptional regulator ADNP links the BAF (SWI/SNF) complexes with autism; Paper presented at the American Journal of Medical Genetics Part C: Seminars in Medical Genetics; 2014.
- Waite KA, Eng C. From developmental disorder to heritable cancer: it's all in the BMP/TGF- β family. *Nature Reviews Genetics.* 2003; 4(10):763–773.
- Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research.* 2010; 38(16):e164–e164. [PubMed: 20601685]
- Ware JS, Samocha KE, Homsy J, Daly MJ. Interpreting de novo variation in human disease using denovolyzeR. *Curr. Protoc. Hum. Gene.* 2015
- Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM, Network CGAR. The cancer genome atlas pan-cancer analysis project. *Nat Genet.* 2013; 45(10):1113–1120. [PubMed: 24071849]
- Yang F, Petsalaki E, Rolland T, Hill DE, Vidal M, Roth FP. Protein domain-level landscape of cancer-type-specific somatic mutations. *PLOS Comput Biol.* 2015; 11(3):e1004147. [PubMed: 25794154]

Figure 1.

Enrichment of germline LGD *de novo* variants in DD patients and controls among candidate cancer driver genes and non-cancer driver genes. Cancer driver genes are grouped based on fraction of LGD somatic mutations among all reported point mutations or small indels in COSMIC. The group with >15% of LGD mutations are likely tumor suppressors. Enrichment values were estimated by comparing observed number of germline *de novo* LGD variants to expectation from background mutation rate (Samocha et al., 2014) in cases or controls. Red bars represent DD cases, blue bars represent controls, error bars represent 95% confident interval. P-values (***) indicates p-value < 0.001) were calculated using Poisson tests with expected value estimated from background mutation rate.

Figure 2.

Class vulnerability of germline *de novo* missense variants in different groups of genes. Class vulnerability is defined as the probability of a variant being associated with the disease (Iossifov et al., 2014). D-mis is defined as missense predicted to be damaging by meta-SVM (Dong et al., 2015). P-values (***) indicates p-value < 0.001; ** indicates p-value < 0.01) were calculated using Binomial tests described in the Materials and Methods section.

Figure 3.

Examples of germline *de novo* missense variants in DD patients superimposed with cancer somatic mutation hotspots. Blue spike lines are somatic missense counts at each amino acid position in cancer. Red dashes indicate predicted hotspot positions by the hidden Markov Model method. Filled diamonds show germline *de novo* variants that are located in somatic hotspots, and hollow diamonds represent germline *de novo* variants that are not located in

somatic hotspots. The following genes are shown: a. *EP300*; b. *ARID2*; c. *DDX3X*; d. *MAP2K1*; e. *PPP2R1A*; f. *CNOT3*; g. *CTNNB1*; h. *SMARCA4*

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Burden of germline *de novo* variants in candidate cancer driver genes (N=1481) comparing to non-driver genes (N=17396). The null hypothesis in Fisher's exact test is that the fraction of germline *de novo* variants of each type located in cancer driver genes is the same between DD cases and controls.

Table 1

Type of <i>de novo</i> variants	Number of <i>de novo</i> variants in 5542 DD cases		Number of <i>de novo</i> variants in 1911 Controls		Fisher's exact	
	Cancer driver genes	Non-cancer driver genes	Cancer driver genes	Non-cancer driver genes	Odds ratio	P-value
LGD	233	671	24	153	2.2	3.3E-04
missense	518	3412	132	1002	1.2	0.19
D-mis	190	766	26	191	1.8	0.006
LGD or D-mis	423	1437	50	344	2	4.5E-06
silent	146	1238	52	427	1	0.86

Number of developmental disorder (DD) candidate risk genes at different FDR values estimated by TADA, and corresponding overlapping cancer driver genes. For each FDR threshold, number of true risk genes is estimated by FDR definition.

Table 2

FDR by TADA	Number of candidate risk genes	Estimated number of true risk genes	Number of candidate risk genes that are cancer drivers	Estimated number of true risk genes that are cancer drivers	Estimated percentage of cancer drivers among true risk genes
0.1	134	120.6	56	54.6	45%
0.2	186	148.8	66	62.2	41%
0.3	269	188.3	74	65.7	34%
0.4	421	252.6	92	74.6	29%
0.5	649	324.4	124	90.5	27%

Enrichment of germline *de novo* missense variants in NDD cases located in cancer somatic missense hotspots.

Table 3

(a) There are significantly more germline <i>de novo</i> missense variants located in hotspots in NDD case comparing to controls.					
Variant type in hotspots	Case	Control	Fold enrichment	p-value	
Missense	95	11	3	0.00013	
D-mis	47	1	16.2	1.8e-5	

(b) Among all germline <i>de novo</i> missense variants in cancer driver genes, the ones in NDD cases are more likely to be located at cancer somatic missense mutation hotspots than the ones in controls.					
Variant type in hotspots	Case		Control		P-value
	hotspots	not hotspots	hotspots	not hotspots	
Missense	95	391	11	102	2.3
D-mis	47	138	1	23	7.8