BMC Medicine

COMMENTARY Open Access

Another tool in the genome-wide association study arsenal: population-based detection of somatic gene conversion

Matthew A Deardorff[1,2*], Jesus Sainz[3,4], Struan FA Grant[1,2,5]

Abstract

The hunt for the genetic contributors to complex disease has used a number of strategies, resulting in the identification of variants associated with many of the common diseases affecting society. However most of the genetic variants detected to date are single nucleotide polymorphisms (SNPs) and copy number variants (CNVs) and fall far short of explaining the full genetic component of any given disease. An as yet untapped genomic mechanism is somatic gene conversion and deletion, which could be complicit in disease risk but has been challenging to detect in genome-wide datasets. In a recent publication in *BMC Medicine* by Kenneth Ross, the author uses existing datasets to look at somatic gene conversion and deletion in human disease. Here, we describe how Ross's recent efforts to detect such occurrences could impact the field going forward.

See research article: http://www.biomedcentral.com/1741-7015/9/12/abstract

Introduction

It is well established that genetic diversity combined with specific environmental exposures contribute to disease susceptibility. However, it has turned out to be challenging to isolate the genes underlying the genetic component conferring susceptibility to most complex disorders. The genetic underpinnings of such traits have remained largely unsolved until relatively recently, where the advent of array-based technologies and large population cohorts have enabled investigators to leverage genetic variation across the entire genome to pinpoint major contributing genetic factors. These discoveries have been primarily driven by genome-wide association studies (GWAS) using single nucleotide polymorphism (SNP) markers, which have revealed compelling evidence, including robust replication, for genetic variants associated with a broad range of phenotypes (see full catalogue at http://www.genome.gov/gwastudies).

These studies have been driven by arrays capable of estimating chromosomal quantitative data as well as SNP genotype status. As such, it has been possible to accurately genotype and rapidly quantify copy number variants (CNVs) [1-3], which have now been strongly implicated in common disorders such as autism [4-7], attention deficit hyperactivity disorder [8], schizophrenia [9-11] and childhood obesity [12].

Nonetheless, these approaches to date have generally only captured a small proportion of the predicted genetic component of various complex traits [13]. It is widely accepted that more extensive meta-analyses and high-throughput sequencing efforts with thousands of DNA samples from affected subjects could lead to further progress. However, these approaches will require large collaborative efforts and robust financial investment, respectively.

While advances are taking place on these fronts, the question remains of whether there are ways that the existing genome-wide SNP datasets could be mined further. After all, many datasets have been deposited in the public domain, most notably those found on dbGaP (http://www.ncbi.nlm.nih.gov/gap). The Wellcome Trust Case Control Consortium (WTCCC) has also made its datasets available to the wider scientific community and has been a key leader in whole genome genetic approaches [14,15].

In a study published this month in *BMC Medicine* [16], Kenneth Ross has made use of the WTCCC genome-wide SNP datasets for 7 common diseases, along with a shared pool of 3,000 controls to ask a focused

* Correspondence: deardorff@email.chop.edu
[1]Division of Human Genetics, The Children's Hospital of Philadelphia Research Institute, Philadelphia, PA, USA
Full list of author information is available at the end of the article

but alternative question. Rather than looking for genetic polymorphisms residing in the germ line, he was interested in uncovering evidence of postzygotic somatic alterations, namely gene conversions and deletions, contributing to the pathogenesis of these diseases. Mitotic gene conversions have been shown to arise as a result of double-strand break repair that uses non-allelic homologous regions [17]. The effects of somatic gene conversion (see Glossary) have been shown to render genes non-functional, impact methylation status and aid the generation of deletions and other copy number variants; indeed, gene conversion has already been implicated in a number of disease settings [17-19].

The reason the approach described is so novel is that detecting these nearly identical recombinants has been technically difficult, due to both technological shortcomings faced by assessing close to identical sequences and difficulties associated with detecting such rare events in the face of a high background 'wild-type' signal.

Ross used the rationale that the genotyping data from most individuals in the WTCCC dataset were derived from blood, representing a population of cells, and that somatic gene conversion in an individual would result in a subtle shift of allele frequency data for an informative SNP. Since these relatively modest alterations can be difficult to detect at the individual level, he assessed whether there were statistical differences in the distribution of the frequency shifts between multiple control and disease populations. To help refine SNPs that were relevant to gene conversion, he used several additional strategies, including limiting analysis to those SNPs associated with regions of homology, and focusing on genotype frequencies that demonstrated unexpected deviation from Hardy-Weinberg equilibrium.

As a consequence of this study design, the author detected multiple instances of putative somatic gene conversion with duplicon identity. Although there is no experimental validation of the detected conversions, the author uses various metrics to assign relative strengths of certainty to the findings. He goes on to speculate on loci impacted by gene conversion and how they may be playing a role in disease.

Although the identified gene conversion is limited to blood, previous data has suggested that significant differences in sister chromatid exchange have been demonstrated in blood from patients with diseases in the WTCCC cohort [15]. Only one of the datasets was from lymphoblastoids and somewhat surprisingly these control samples did not show large differences from the blood genotyped controls.

This approach provides a new complementary methodology to detect gene conversion for regions where the CNV status has been previously characterized. This technique will, however, be somewhat more limited for variability still to be defined in specific individuals; indeed currently available genomic sequencing data suggests that such variability is extensive.

With these caveats in mind, and the fact that the analyses were limited to considering homologous regions, it is clear that this current study is primarily hypothesis forming, with various loci presented as potentially playing a role in disease risk. Nonetheless these hypotheses are testable, and the gene conversions identified by Ross can be tested in future datasets from DNA derived directly from target tissues or blood from other replication cohorts to further clarify their roles in these diseases. Once replicated, the field can move forward with greater certainty that perhaps at least one these gene conversion loci are contributing to disease risk and functional studies can be carried out to determine mode of action.

Glossary

Somatic gene conversion

This concept defines the process by which DNA sequence information is transferred in a non-reciprocal process from one genomic region to another region of the genome, altering its sequence. The transfer of genomic information is due to base mismatch repair during the recombination in somatic division

Duplicons

These are duplicated genomic segments, also known as segmental duplications. These elements are large genomic segments of recent origin and nearly identical sequence present as low copy repeats. The length of duplicons can vary from 1 kb to hundreds of kb and have a high level of sequence identity (>90%)

Author details
[1]Division of Human Genetics, The Children's Hospital of Philadelphia Research Institute, Philadelphia, PA, USA. [2]Department of Pediatrics, The University of Pennsylvania School of Medicine, Philadelphia, PA, USA. [3]Institute of Biomedicine and Biotechnology of Cantabria (IBBTEC), Faculty of Medicine, University of Cantabria, Santander, Spain. [4]Spanish National Research Council (CSIC), Madrid, Spain. [5]Center for Applied Genomics, The Children's Hospital of Philadelphia Research Institute, Philadelphia, PA, USA.

Authors' contributions
All authors contributed equally to this work.

Competing interests
The authors declare that they have no competing interests.

Received: 25 November 2010 Accepted: 3 February 2011
Published: 3 February 2011

References

1. Reich D, Patterson N, De Jager PL, McDonald GJ, Waliszewska A, Tandon A, Lincoln RR, DeLoa C, Fruhan SA, Cabre P, Bera O, Semana G, Kelly MA, Francis DA, Ardlie K, Khan O, Cree BA, Hauser SL, Oksenberg JR, Hafler DA: **A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility.** *Nat Genet* 2005, **37**:1113-1118.
2. Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL: **Whole-genome genotyping with the single-base extension assay.** *Nat Methods* 2006, **3**:31-33.
3. Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS: **A genome-wide scalable SNP genotyping assay using microarray technology.** *Nat Genet* 2005, **37**:549-554.

4. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimäki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, *et al*: **Strong association of *de novo* copy number mutations with autism.** *Science* 2007, **316**:445-449.
5. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y, Thiruvahindrapduram B, Fiebig A, Schreiber S, Friedman J, Ketelaars CE, Vos YJ, Ficicioglu C, Kirkpatrick S, Nicolson R, Sloman L, Summers A, Gibbons CA, Teebi A, Chitayat D, Weksberg R, Thompson A, Vardy C, Crosbie V, Luscombe S, Baatjes R, *et al*: **Structural variation of chromosomes in autism spectrum disorder.** *Am J Hum Genet* 2008, **82**:477-488.
6. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T, Platt OS, Ruderfer DM, Walsh CA, Altshuler D, Chakravarti A, Tanzi RE, Stefansson K, Santangelo SL, Gusella JF, Sklar P, Wu BL, Daly MJ, Autism Consortium: **Association between microdeletion and microduplication at 16p11.2 and autism.** *N Engl J Med* 2008, **358**:667-675.
7. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, Imielinski M, Frackelton EC, Reichert J, Crawford EL, Munson J, Sleiman PM, Chiavacci R, Annaiah K, Thomas K, Hou C, Glaberson W, Flory J, Otieno F, Garris M, Soorya L, Klei L, Piven J, Meyer KJ, Anagnostou E, Sakurai T, *et al*: **Autism genome-wide copy number variation reveals ubiquitin and neuronal genes.** *Nature* 2009, **459**:569-573.
8. Elia J, Gai X, Xie HM, Perin JC, Geiger E, Glessner JT, D'Arcy M, deBerardinis R, Frackelton E, Kim C, Lantieri F, Muganga BM, Wang L, Takeda T, Rappaport EF, Grant SF, Berrettini W, Devoto M, Shaikh TH, Hakonarson H, White PS: **Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes.** *Mol Psychiatry* **15**:637-646.
9. Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Möller HJ, Hartmann A, *et al*: **Large recurrent microdeletions associated with schizophrenia.** *Nature* 2008, **455**:232-236.
10. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, *et al*: **Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia.** *Science* 2008, **320**:539-543.
11. Glessner JT, Reilly MP, Kim CE, Takahashi N, Albano A, Hou C, Bradfield JP, Zhang H, Sleiman PM, Flory JH, Imielinski M, Frackelton EC, Chiavacci R, Thomas KA, Garris M, Otieno FG, Davidson M, Weiser M, Reichenberg A, Davis KL, Friedman JI, Cappola TP, Margulies KB, Rader DJ, Grant SF, Buxbaum JD, Gur RE, Hakonarson H: **Strong synaptic transmission impact by copy number variations in schizophrenia.** *Proc Natl Acad Sci USA* 2010, **107**:10584-10589.
12. Glessner JT, Bradfield JP, Wang K, Takahashi N, Zhang H, Sleiman PM, Mentch FD, Kim CE, Hou C, Thomas KA, Garris ML, Deliard S, Frackelton EC, Otieno FG, Zhao J, Chiavacci RM, Li M, Buxbaum JD, Berkowitz RI, Hakonarson H, Grant SF: **A genome-wide study reveals copy number variants exclusive to childhood obesity cases.** *Am J Human Genet* 2010, **87**:661-666.
13. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461**:747-753.
14. Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447**:661-678.
15. The Wellcome Trust Case Control Consortium: **Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls.** *Nature* 2010, **464**:713-720.
16. Ross KA: **Evidence for somatic gene conversion and deletion in bipolar disorder, Crohn's disease, coronary artery disease, hypertension, rheumatoid arthritis, type-1 diabetes, and type-2 diabetes.** *BMC Med* 2011, **9**:12.
17. Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP: **Gene conversion: mechanisms, evolution and human disease.** *Nat Rev* 2007, **8**:762-775.
18. Johnson RD, Jasin M: **Double-strand-break-induced homologous recombination in mammalian cells.** *Biochem Soc Transact* 2001, **29**:196-201.
19. Lagerstedt K, Karsten SL, Carlberg BM, Kleijer WJ, Tonnesen T, Pettersson U, Bondeson ML: **Double-strand breaks may initiate the inversion mutation causing the Hunter syndrome.** *Hum Mol Genet* 1997, **6**:627-633.

Pre-publication history
The pre-publication history for this paper can be accessed here:
http://www.biomedcentral.com/1741-7015/9/13/prepub

doi:10.1186/1741-7015-9-13
Cite this article as: Deardorff *et al.*: Another tool in the genome-wide association study arsenal: population-based detection of somatic gene conversion. *BMC Medicine* 2011 **9**:13.