GWATCH: A Web Platform for Automated Gene Association Discovery Analysis

Anton Svitin[1]*†, Sergey Malov[1,2]*, Nikolay Cherkasov[1]*, Paul Geerts[3], Mikhail Rotkevich[1], Pavel Dobrynin[1], Andrey Shevchenko[1], Li Guan[1], Jennifer Troyer[4], Sher Hendrickson-Lambert[5], Holli Hutcheson Dilks[6], Taras K. Oleksyk[7], Sharyne Donfield[8], Edward Gomperts[9], Douglas A. Jabs[10], Mark Van Natta[11], Richard Harrigan[12,13], Zabrina L. Brumme[14], and Stephen J. O'Brien[1,15]†

[1]Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, 41 Sredniy Prospekt, St. Petersburg, Russia 199004.

[2]Department of Mathematics, St.-Petersburg Electrotechnical University, 5 Prof. Popova St., St. Petersburg, Russia 197376.

[3]Scientific Data Visualization Consultant, Turner, ACT, Australia.

[4]Genetics and Genomics Group, Advanced Technology Program, SAIC-Frederick, National Cancer Institute, Frederick, MD.

[5]Department of Evolutionary Biology, Shepherd University, Shepherdstown, WV 25443.

[6]Vanderbilt Technologies for Advanced Genomics, Office of Research, Vanderbilt University Medical Center, Nashville, TN.

[7]Biology Department, University of Puerto Rico, Mayaguez, Puerto Rico 00680, USA.

[8]Department of Biostatistics, Rho, Inc., Chapel Hill, NC, USA.

[9]Division of Hematology-Oncology, Children's Hospital of Los Angeles, Los Angeles, CA, USA.

[10]Departments of Ophthalmology and Medicine, Icahn School of Medicine at Mount Sinai, New York, NY.

[11]Department of Epidemiology, The Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD.

[12]British Columbia Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada V6Z 1Y6.

[13]Division of AIDS, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada V6T 1Z3.

[14]Faculty of Health Sciences, Simon Fraser University, Burnaby, B.C., Canada.

[15]Oceanographic Center, Nova Southeastern University, Ft. Lauderdale, FL.

†Correspondence to: Anton Svitin anton.svitin@gmail.com and S.J. O'Brien lgdchief@gmail.com.

*AS, SM and NC contributed equally to this work.

As genome-wide sequence analyses for complex diseases expand, web-tools promoting analytical validation of associations are required. GWATCH (Genome-Wide Association Track Chromosome Highway) is a dynamic genome browser that automates: 1.) Rapid gene association discovery analysis of large genome-wide data sets; 2.) Expanded visual display for SNP-indel-CNV gene associations including: Manhattan plots, 2D and 3D snapshots of gene regions, and a dynamic genome spanning browser; 3.) Real time validation/replication of suggested candidate or putative genes, reducing Bonferroni penalties; 4.) Open unabridged data release, respecting privacy constraints and facilitating pubic comparative meta-analysis. We illustrate GWATCH with a large proven HIV-AIDS gene association dataset, though any disease association dataset can be uploaded.

Annotations of human genome variation have identified some 12 million single nucleotide polymorphisms (SNPs), which offer the promise of connecting nucleotide and structural variation to hereditary traits (Wellcome Trust Case Control Consortium 2007; International HapMap 3 Consortium et al. 2010; 1000 Genomes Project Consortium et al. 2012). Genotyping arrays that resolve millions of common SNPs have enabled >1000 Genome-Wide Association Studies (GWAS) to discover principal genetic determinants of complex multifactorial human diseases (Hindorff,L.A., MacArthur,J., Morales,J., Junkins,H.A., Hall,P.N., Klemm,A.K. and Manolio,T.A. A catalog of published genome-wide association studies, available at: www.genome.gov/gwastudies.Related citations; Hindorff et al. 2009). Today whole genome sequence association has extended the prospects for personalized genomic medicine, capturing rare variants, copy number variation, indels, epistatic and epigenetic interactions in hopes of achieving individualized genomic assessment, diagnostics, and therapy of complex maladies (Cirulli and Goldstein 2010; Jiang et al. 2013; Kilpivaara and Aaltonen 2013; Wade et al. 2013). Indeed, the application of whole genome sequencing to hundreds of vertebrate species is now allowing association studies of model species (mice, cats, pigs etc.) plus any species from which DNA variants are annotated (Genome 10K Community of Scientists 2009).

A challenge to genetic epidemiology involves disentangling true functional associations that fall below genome-wide significance threshold from statistical artifacts (Conneely and Boehnke 2007; Dudbridge and Gusnanto 2008; McCarthy et al. 2008; Moskvina and Schmidt 2008; Bushman et al. 2009; Goldstein 2009; Ioannidis et al. 2009; Johnson et al. 2010; O'Brien and Hendrickson 2013; "Best practices in GWAS", *Genome Technology* Supplemental report 2009, available at http://www.genomeweb.com/node/917734). Researchers agree that open access sharing of GWAS data would allow rapid replication and inspection approaches to this question (Johnson and O'Donnell 2009; Hayden 2013). However, for many cohorts, participants have not consented to open access of personal data. Since patient anonymization is virtually impossible with genetic data, an open posting of patients' genotype and clinical data ethically conflicts with promised individual privacy (Greely 2007; O'Brien 2009; Gymrek et al. 2013). **GWATCH (**Genome-Wide Association Track Chromosome

Highway; gen-watch.org) overcomes this issue through an organized open release of unabridged SNP-test association results from GWAS and whole genome sequencing (WGS) association studies while sequestering individuals' personal data.

GWATCH is a dynamic genome browser that automates primary analysis and displays its results: p-values and Quantitative Association Statistic (QAS, a general term for statistics explaining direction and strength of associations: odds ratio, relative hazard and ez2-transformed correlation coefficient; section 2.2 of Materials and Methods in Supplementary Information) from multiple association tests performed for one or more cohorts in a GWAS or WGS study as a visual array ordered by SNP chromosomal position. **GWATCH** offers "features" that allow automated analysis and visualization of multiple test outcomes, rapid discovery, replication and data-release of unabridged association results (Table 1). We illustrate the utility, interpretation, and navigation of **GWATCH** using a GWAS meta-analysis carried out with 5922 study participants enrolled in eight prospective HIV-AIDS cohorts, searching for AIDS Restriction Genes (Dean et al. 1996; Chinn et al. 2010; Hendrickson et al. 2010; Herbeck et al. 2010; Troyer et al. 2011; O'Brien and Hendrickson 2013).

Typical input of a GWAS analysis includes a large unabridged **Data-Table** listing p-values and QASs across multiple SNP association tests performed for a list of ~1-10,000,000 ordered SNPs (Supplementary Table 1). **GWATCH** displays the Data-Table, association tests and their results in various forms: Manhattan plots for each test, 2D and 3D snapshots of test results for chromosome regions of "hits", and a dynamic chromosome browser that illustrates significant p-values and QASs (Supplementary Figures 1-4). The browser provides a dynamic traverse along a human chromosome producing a "bird's eye" view of the strong SNP associations that rise above the chromosome highway surface. The idea is to visualize association results across a gene region (e. g. one that may include a highly significant SNP association) for all the tests performed and for all the neighboring, potentially proxy SNPs (i.e. SNPs which track the neighboring causal, disease-causing SNP due to the linkage disequilibrium - LD) for the same tests.

Top hits are ranked based upon extreme p-values, QASs, or "density" of composite p-value peaks (representing proxy SNPs in linkage disequilibrium and multiple non-independent association tests). A multi-page "TRAX Report" produces unabridged curves, tables and appropriate statistics for any SNP-variant (Supplementary Figures 5-6). **GWATCH** automates the computation and visualization of results allowing instant replication of putative discoveries suggested by outside cohort studies or functional experiments.

GWATCH provides a dynamic visual journey, similar to driving a video game along human chromosomes to view patterns of GWAS- or WGS-based variant association with any complex disease. It is appealing, intuitive, and accessible to non-experts. **GWATCH** will import new data from any disease gene association study with multiple disease stages or analytical strategies. The wide breadth of test associations displayed is particularly suited to complex disease cohorts with detailed clinical parameters over distinct disease stages. Further, **GWATCH** facilitates rapid replication of gene discoveries from independent cohort studies by simply keying in the putative gene region and inspecting the many test results of the posted dataset. Since replication screens are hypothesis-driven, they avoid the stringent multiple test correction penalties of a GWAS/WGS ($p<10^{-8}$). Further, different cohort studies can be compared directly or combined to build meta-analyses.

Should many cohort investigators release their unabridged results, then association discoveries will be replicated in a rapid, open and productive manner, allowing for meta-analyses as have been proposed for HIV-AIDS and other complex diseases (Johnson and O'Donnell 2009; Hayden 2013; McLaren et al. 2013). Unlike other methods of data sharing, the **GWATCH** approach avoids any violation of patient privacy, HIPPA concerns, or informed consent constraints, since the primary clinical and genotype data remain confidential while the derivative results (p-values, QASs, plots) of multiple analytical approaches are released. Our hope would be to expand discovery and replication opportunities in biomedical, comparative and evolutionary genomic research (e.g. genome scans for selection signatures or population divergence) assuring maximum benefit of data sharing while protecting patients who prefer privacy, but wish to see their volunteerism fulfilled.

GWATCH is available online at gen-watch.org.

Supplementary Material including Materials and Methods, Supplementary Figures 1-6 and Supplementary Tables 1-10 is available online.

Acknowledgments

This work was supported in part by Russian Ministry of Science Mega-grant no.11.G34.31.0068; Stephen J. O'Brien, Principal Investigator. The Hemophilia Growth and Development Study is funded by the National Institutes of Health, National Institute of Child Health and Human Development, R01-HD-41224.

References:

1. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56-65.
2. Bushman FD, Malani N, Fernandes J, D'Orso I, Cagney G, Diamond TL, Zhou H, Hazuda DJ, Espeseth AS, König R, et al. 2009. Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathog.* 5:e1000437.
3. Chinn LW, Tang M, Kessing BD, Lautenberger JA, Troyer JL, Malasky MJ, McIntosh C, Kirk GD, Wolinsky SM, Buchbinder SP, et al. 2010. Genetic associations of variants in genes encoding HIV-dependency factors required for HIV-1 infection. *J Infect Dis.* 202:1836-1845.
4. Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet.* 11:415-425.
5. Conneely KN, Boehnke M. 2007. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *Am J Hum Genet.* 81:1158-1168.
6. Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets R, Goedert JJ, Buchbinder SP, Vittinghoff E, Gomperts E, et al. 1996. Genetic restriction of HIV-1

infection and progression to AIDS by a deletion allele of the CKR5 structural gene. *Science* 273:1856-1862.

7. Dudbridge F, Gusnanto A. 2008. Estimation of significance thresholds for genome wide association scans. *Genet Epidemiol.* 32:227–234.
8. Genome 10K Community of Scientists. 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered.* 100:659-674.
9. Goldstein DB. 2009. Common genetic variation and human traits. *N Engl J Med.* 360:1696-1698.
10. Greely HT. 2007. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. *Annu Rev Genomics Hum Genet.* 8:343-364.
11. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. 2013. Identifying personal genomes by surname inference. *Science* 339:321-324.
12. Hayden EC. 2013. Geneticists push for global data-sharing. *Nature* **498:** 16-17.
13. Hendrickson SL, Lautenberger JA, Chinn LW, Malasky M, Sezgin E, Kingsley LA, Goedert JJ, Kirk GD, Gomperts ED, Buchbinder SP, et al. 2010. Genetic variants in nuclear-encoded mitochondrial genes influence AIDS progression. *PLoS One* 5:e12862.
14. Herbeck JT, Gottlieb GS, Winkler CA, Nelson GW, An P, Maust BS, Wong KG, Troyer JL, Goedert JJ, Kessing BD, et al. 2010. Multistage genomewide association study identifies a locus at 1q41 associated with rate of HIV-1 disease progression to clinical AIDS. *J Infect Dis.* 201:618-626.
15. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A.* 106:9362-9367.
16. International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52-58.

17. Ioannidis JP, Thomas G, Daly MJ. 2009. Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet.* 10:318-329.

18. Jiang YH, Yuen RK, Jin X, Wang M, Chen N, Wu X, Ju J, Mei J, Shi Y, He M, et al. 2013. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am J Hum Genet.* 93:249-263.

19. Johnson AD, O'Donnell CJ. 2009. An open access database of genome-wide association results. *BMC Med Genet.* 10:6.

20. Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, O'Brien SJ. 2010. Accounting for multiple comparisons in a genome wide association study (GWAS). *BMC Genomics* 11:724.

21. Kilpivaara O, Aaltonen LA. 2013. Diagnostic cancer genome sequencing and the contribution of germline variants. *Science* 339:1559-1562.

22. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet.* 9:356-369.

23. McLaren PJ, Coulonges C, Ripke S, van den Berg L, Buchbinder S, Carrington M, Cossarizza A, Dalmau J, Deeks SG, Delaneau O, et al. 2013. Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls. *PLOS Pathog.* 9:e1003515.

24. Moskvina V, Schmidt KM. 2008. On multiple-testing correction in genome-wide association studies. *Genet Epidemiol.* 32:567-573.

25. O'Brien SJ. 2009. Stewardship of human biospecimens, DNA, genotype, and clinical data in the GWAS era. *Annu Rev Genomics Hum Genet.* 10:193–209.

26. O'Brien SJ, Hendrickson S. 2013. Host genomic influences on HIV/AIDS. *Genome Biol.* 14:201.

27. Troyer JL, Nelson GW, Lautenberger JA, Chinn L, McIntosh C, Johnson RC, Sezgin E, Kessing B, Malasky M, Hendrickson SL, et al. 2011. Genome-wide association study implicates PARD3B-based AIDS restriction. *J Infect Dis.* 203:1491-1502.

28. Wade CH, Tarini BA, Wilfond BS. 2013. Growing up in the genomic era: implications of whole-genome sequencing for children, families, and pediatric practice. *Annu Rev Genomics Hum Genet.* 14:535-555.

29. Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661-678.

Table 1. Display Feature Components of GWATCH.

Features displayed	Illustration
1. **Unabridged Data Table** of SNP chromosome coordinates, MAF[a], p-value and QAS[b] for each SNP for each test	Supp. Table 1
2. **Association Tests List** and **Manhattan Plots** for each test across all SNPs	Supp. Figure 1
3. **SNAPSHOTS** of SNP-test results in a chromosome region:	
1. **2D Heat Plot Snapshot** illustrating p-values in any selected chromosome region	Supp. Figure 2
2. **3D Checkerboard Plot Snapshot** illustrating p-values and QAS[b] in any selected chromosome region	Supp. Figure 3
3. **LD-polarized 3D Checkerboard Snapshot** illustrating p-values and QAS[b] in any selected chromosome region	Supp. Figure 4
4. **Dynamic Highway View by Chromosome Browser** illustrating p-values and QAS[b]	GWATCH website
5. **Top association hits:**	
1. **Top hits** based on **ranked -log p-value**	Supp. Table 2
2. **Top hits** based on **ranked QAS**[b]	Supp. Table 2
3. **Top hits** based on **ranked Density of -log p-value** within a SNP genomic region	Supp. Table 2
6. **TRAX feature:**	
1. **TRAX PAGE** – two-page graphic summary illustrating p-values and QAS[b] for one selected SNP	Supp. Figure 5
2. **TRAX REPORT** – eleven-page analysis summary with graphs, curves and tables for all association tests for one selected SNP	Supp. Figure 6

Abbreviations: [a]MAF - minor allele frequency; [b]QAS - quantitative association statistic (OR, RH, ez2-transformed correlation coefficient).