




RESEARCH ARTICLE

Low budget analysis of Direct-To-Consumer genomic testing familial data [version 1; peer review: 2 approved]

Gustavo Glusman¹, Mike Carriaso², Rafael Jimenez³, Daniel Swan⁴, Bastian Greshake⁵, Jong Bhak⁶, Darren W Logan⁷, Manuel Corpas ⁸

¹Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109-5234, USA

²River Road Bio LLC, Potomac, MD 20854-3976, USA

³European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

⁴Oxford Gene Technology, Begbroke Science Park, Begbroke, Oxfordshire, OX5 1PF, UK

⁵Molecular Ecology Group, Biodiversity and Climate Research Centre, Frankfurt am Main, Senckenberganlage 25, D-60325, Germany

⁶Theragen BiO Institute, TheragenEtex Inc, AICT building, Lui-dong, Youngtong-gu, Suwon 443-370, South Korea

⁷Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

⁸The Genome Analysis Centre, Norwich Research Park, Norwich, NR4 7UH, UK



v1 First published: 16 Jul 2012, 1:3 (<https://doi.org/10.12688/f1000research.1-3.v1>)
 Latest published: 16 Jul 2012, 1:3 (<https://doi.org/10.12688/f1000research.1-3.v1>)

Abstract

Direct-to-consumer (DTC) genetic testing is a recent commercial endeavor that allows the general public to access personal genomic data. The growing availability of personal genomic data has in turn stimulated the development of non-commercial tools for DTC data analysis. Despite this new wealth of public resources, no systematic research has been carried out to assess these tools for interpretation of DTC data. Here, we provide an initial analysis benchmark in the context of a whole family, using single nucleotide polymorphism (SNP) data. Five blood-related DTC SNP chip data tests were analyzed in conjunction with one whole exome sequence. We report findings related to genomic similarity between individuals, genetic risks and an overall assessment of data quality; thus providing an evaluation of the current potential of public domain analysis tools for personal genomics. We envisage that as the use of personal genome tests spreads to the general population, publicly available tools will have a more prominent role in the interpretation of genomic data in the context of health risks and ancestry.

Open Peer Review

Reviewer Status  

	Invited Reviewers	
	1	2
version 1 published 16 Jul 2012	 report	 report

- Peter N. Robinson**, Institute for Medical Genetics, Universitätsklinikum Charité, Berlin, Germany
- Christian Gilissen**, Radboud University, Nijmegen Medical Center, Nijmegen, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Manuel Corpas (mc@manuelcorpas.com)

Competing interests: Michael Cariaso is the developer of Promethease and SNPedia. Some aspects of the Promethease analysis shown here cost \$2 per individual.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2012 Glusman G *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

How to cite this article: Glusman G, Cariaso M, Jimenez R *et al.* **Low budget analysis of Direct-To-Consumer genomic testing familial data [version 1; peer review: 2 approved]** F1000Research 2012, 1:3 (<https://doi.org/10.12688/f1000research.1-3.v1>)

First published: 16 Jul 2012, 1:3 (<https://doi.org/10.12688/f1000research.1-3.v1>)

Introduction

Direct-to-Consumer (DTC) genetic testing is a relatively new commercial endeavor offering access to personal genomic tests to the general public. Individuals wishing to learn about their genomes today enjoy a range of options. DTC providers typically offer chip-based genotyping of genome-wide markers, currently in the range of hundreds of thousands to a million single-nucleotide polymorphisms (SNPs). This current wealth of personal genomic data is likely to grow at an increasing pace as DNA sequencing become ubiquitous in personal genome testing. Genome sequencing allows elucidation of not just SNPs, but copy number variants (CNVs), insertions, inversions and many other genomic features currently underrepresented in personal genome analyses. Yet personal SNP data has proven to be a valuable resource for making personalized inferences about the risk of developing medical conditions, the probability of having certain phenotypic traits, and one's likely ancestral origins^{1,2}. Taking advantage of the growing body of statistical associations accumulated in the scientific literature, DTC providers have been able to offer personalized genomic 'reports' that present accessible scientific information of relevance to their customers' observed genotypes. It is precisely in these genomic annotations where customers realize the value of their DTC product purchase.

A feature of DTC genomic test interpretation is that, being a commercial product, genomic annotations and analysis tools are proprietary and not freely available to the research community. This has motivated the parallel development of public resources and low cost genotype analysis tools. SNPedia³ is a wiki-styled resource that collects and annotates SNPs from the scientific literature and provides tools with which to associate these annotations to those observed in DTC genomic tests. openSNP⁴ is a public resource that collects genotypes from people willing to share them, allows annotation of phenotypes, and the search of occurrences of a particular SNP in scientific publications using Mendeley⁵. Although all resources, public or commercial, are limited by the reliability of the data available for any given marker, public and low cost resources have the potential of engaging community wide efforts ('crowd-sourcing') to an extent to which closed commercial applications cannot.

We decided to explore the extent to which phenotype inference and genotype analysis can be carried out solely using existing public or very low cost resources. This is motivated by our belief that no DTC company will ultimately be able to match the rapid pace of genomic data accumulation and annotation that the research community is producing. Apart from SNPedia or openSNP, a wiki-based model, perhaps integrated with existing genetics resources such as the Gene Wiki⁶ or Gene Wiki+⁷, may offer a good solution for rapid, accurate and comprehensive community annotation of personal genomic data.

In this paper we carry out a systematic analysis of DTC genomic data from a family of five blood relatives using mostly public annotations and tools. We present a) our findings related to the quality of the data, b) a comparison of the similarity between members of the family and an undisclosed individual of a different ethnic background and c) phenotype inferences as described by SNPedia trait annotations. We also incorporate analysis of DTC exome sequence data to supplement the genotype findings of one individual. We thus offer a pioneering methodical study of a whole family analyzed using only

DTC data. Since comparable data could, in principle, be bought by any individual, this study benchmarks the personal genomic analyses available to non-experts using open, web-based tools.

Results

We analyzed DTC genomic data from a family of five of self-reported European ancestry, two males and three females across two generations (Figure 1).

The family has lived in the southern-most region of Western Europe (Andalusia, Spain) for at least 4 generations. Principal component analyses of ancestry informative markers confirm tight parental clustering with Southern European populations (Figure 2). We thus expect their ethnic background to be relatively homogeneous. The CEU HapMap ethnic group⁸ was taken as the reference genotype for SNPedia phenotype predictions. Two kinds of SNP chip were used in this analysis, the 23andMe⁹ versions 2 and 3. All family members except 'Son' (denoted with red diamond in Figure 1) were tested with version 3. Son was tested using version 2 and whole exome sequencing.

Family SNP chip analysis

Whole family genotypes provide an additional genetic context that individual data analyses cannot offer, enabling enhanced error correction and inheritance state analysis¹⁰. We found that data downloaded from 23andMe at different times may vary, probably as a consequence of changes in genotype-calling algorithms. To ensure consistency in our analyses, we downloaded the most recent data, for all family members, on May 30th 2012.

23andMe SNP chip genotype data

5 Data Files

<http://dx.doi.org/10.6084/m9.figshare.92682>

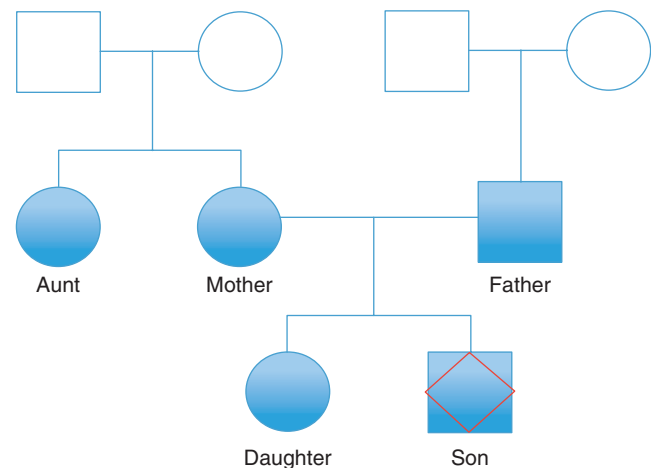


Figure 1. Family tree analyzed using DTC genotyping services. Squares and circles denote male and female respectively. Filled shapes represent those for which genome data is available. 'Son', the individual whose exome was sequenced, is denoted with a red diamond. Other family members include Father, Mother, Daughter and Aunt, who is Mother's sister.

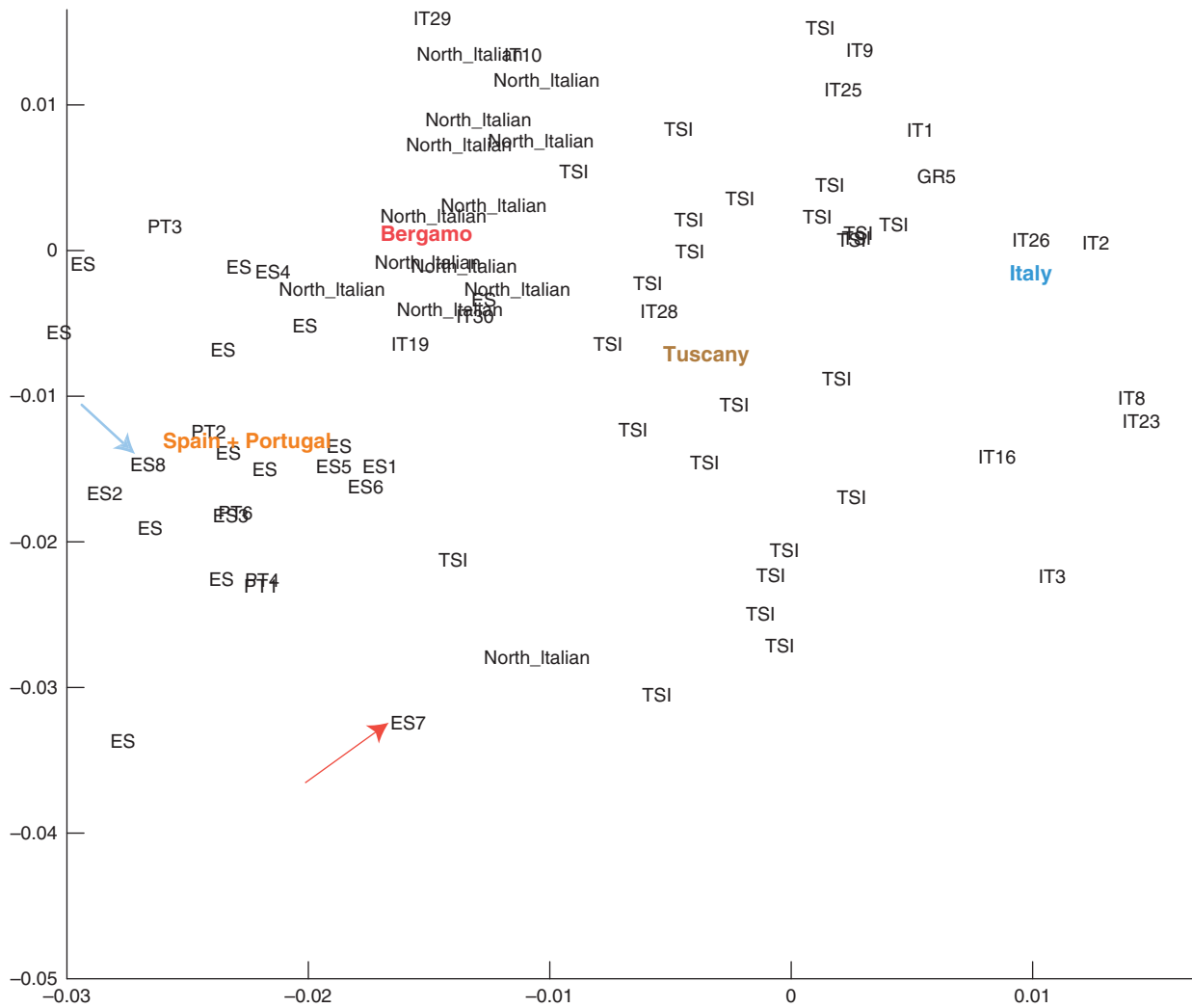


Figure 2. Admixture analysis of individuals from Southern Europe from the Eurogenes Genetic Ancestry Project. Mother (ES7) is denoted by a red arrow and Father (ES8) by a blue arrow. Mother and Father are the only family individuals included here as they have the most divergent genotypes within the family.

In order to facilitate comparison between different genotypes, we excluded non-autosomal SNP data (chromosomes X, Y and MT). All v3 chips had a total of 930,342 autosomal SNPs; the v2 SNP chip had 556,694. The reported ‘no call’ rate (shown as ‘-’ in the downloaded data) varied slightly for each individual (Table 1), but overall, when genotypes are expressed as percentages of the total,

some differences are observed for Son (v2) when compared to all other v3 individuals (Figure 3). V3 individuals show a very similar distribution of genotypes.

Calculation of error rates

23andMe reports a 98% or greater call rate¹¹, meaning that the chip can provide accurate data for more than 98% of those variants in any particular person. When an allele variant present in heterozygous state is “undercalled” (not observed), the locus may be reported as being homozygous for the other variant, leading to missed heterozygosity. Such sites may significantly impact the disease risks predicted for the individual. Under the simplifying assumption of uniform undercall probability, we estimated the number of heterozygous sites mistakenly reported as homozygous (Table 2). This means that for Son, 1 in every 400 sites is mistakenly called. For Father, 1 in every 200. Next, we analyzed Mendelian Inheritance Errors (MIEs). If at one site a reported genotype is ‘CC’ but the genotypes for both parents is ‘TT’, one possible explanation

Table 1. Summary of reported no-call rates for all 23andMe chips included in this study.

Family member	No call rate	Chip version
Mother	0.24%	v3
Father	0.21%	v3
Daughter	0.19%	v3
Aunt	0.17%	v3
Son	0.16%	v2

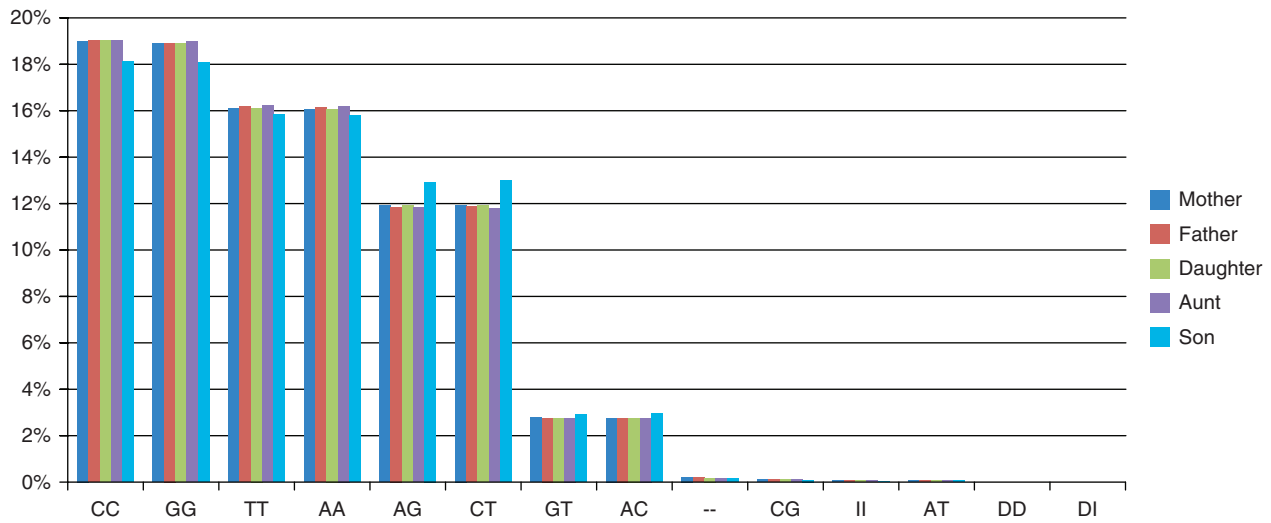


Figure 3. A distribution of all the different occurring genotypes as a percentage of the total for all individuals is shown. For the purposes of unbiased comparison, only autosome data is included. I and D indicate insertion and deletion, respectively. Son's percentages (v2) show slight differences to all other v3 individuals whose genotype proportions are more similar.

Table 2. Number of heterozygous sites mistakenly reported as homozygous (based on the undercall rate, in autosomes).

Member	Undercall	Heterozygous to Homozygous
Son	0.25%	661
Daughter	0.53%	2007
Mother	0.60%	2269
Father	0.50%	2010
Aunt	0.51%	1905

is that one of the parents is actually heterozygous 'CT' but was undercalled as 'CC', and likewise the son is heterozygous 'CT' but was undercalled 'TT'. Given 5 people, there are 10 possible pairwise relations. Four of these represent direct parent/offspring relations, for which discrepancies can be counted as MIEs (Table 3).

One SNP in the Daughter (chr4:7, 9957, 622) is in disagreement with both parents: Father=CC, Mother=CC, Daughter=TT.

Table 3. Mendelian Inheritance errors as estimated by direct parent/offspring relations.

Relation	MIEs
Son/Father:	36
Son/Mother:	24
Daughter/Father:	108
Daughter/Mother:	129

For the remaining 6 relationships, only part of the genome is expected to be identical by descent (IBD). Fully incompatible sites can be numbered as "pseudo-MIEs" (Table 4).

The reduced numbers of MIEs and pseudo-MIEs between the Son and all other family members were due to the lower total number of SNPs assayed for the Son. The Daughter has more MIEs relative to the Mother than relative to the Father. This may be due to having inherited a few deleted segments leading to a hemizygous state: hemizygous sites are reported as homozygous for the allele present, leading to an accumulation of apparent MIE sites. For example, at chr2:41093584 (rs12465519) (Table 5). The ISCA analysis explains how we inferred a deletion from observed discordant genotypes.

A deletion inferred from mismatching genotype data

1 Data File

<http://dx.doi.org/10.6084/m9.figshare.92821>

Table 4. Incompatible sites between remaining relationships, identified as pseudo-MIEs.

Relation	Pseudo-MIEs
Son/Daughter [siblings]:	7,777
Mother/Aunt [siblings]:	15,401
Son/Aunt:	17,390
Daughter/Aunt:	30,215
Father/Mother [unrelated]:	53,937
Father/Aunt [unrelated]:	54,522

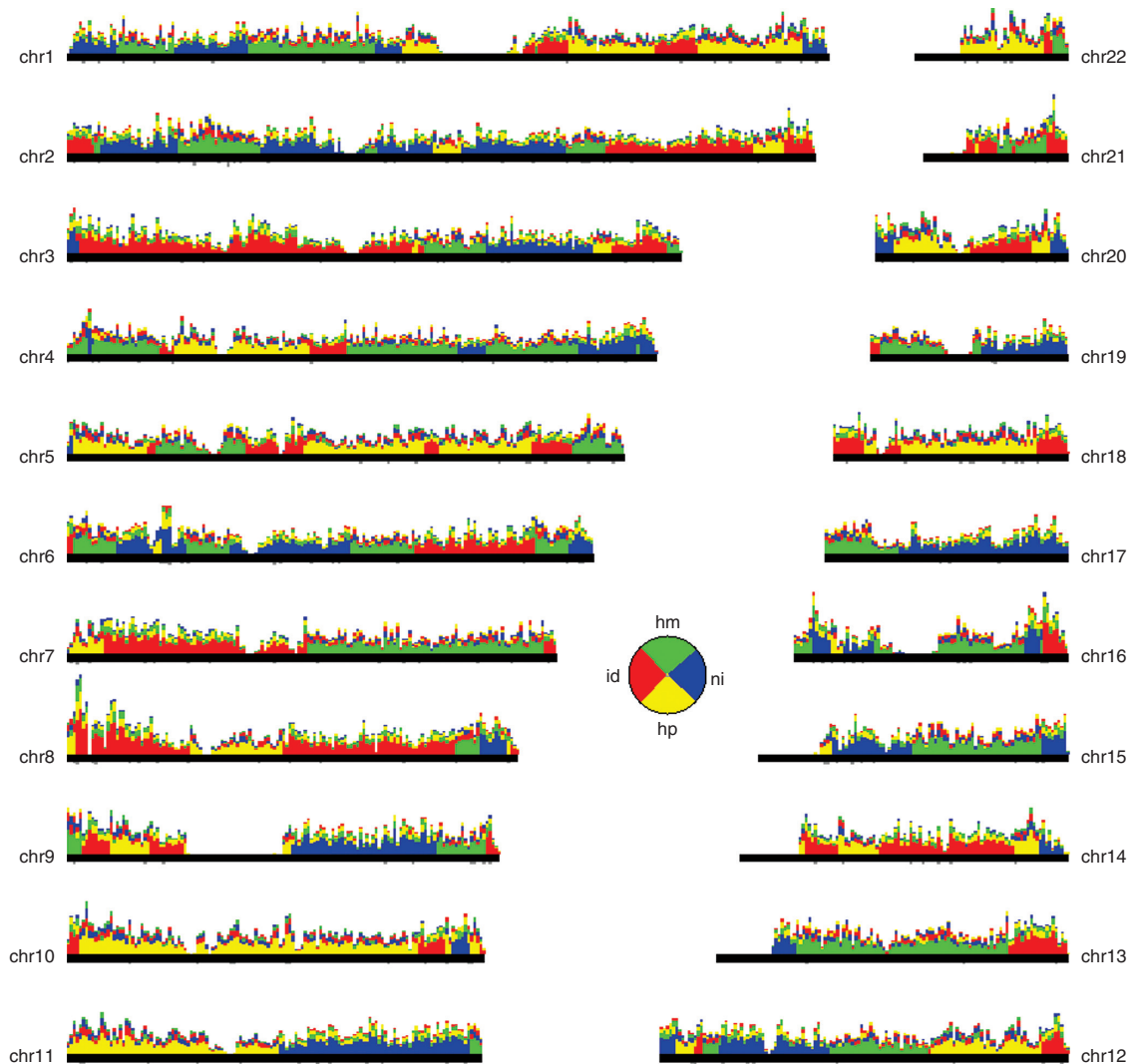


Figure 4. Inheritance State Consistency Analysis (ISCA) plot for the Father-Mother-Son-Daughter quartet, depicting for each autosome the number of informative SNPs supporting each of the four possible inheritance states: “identical” (“id”, red), “haploidentical maternal” (“hm”, green), “haploidentical paternal” (“hp”, yellow) and “nonidentical” (“ni”, blue). SNPs consistent with two inheritance states contribute 0.5 weight to each. SNP counts are binned in non-overlapping 1 Mb windows; within each window, the four inheritance states are sorted by decreasing level of support. Regions without support typically overlap centromeric repeats and heterochromatic regions. Pie chart inset: fraction of the genome observed in each inheritance state.

Father/Daughter vs. Mother/Daughter similarity was found significant (p -value = $1.731e-05$). Son however, exhibited 83.7% identity to Mother and 83.8% to Father and his tested SNPs were not found to be significantly different to either of them (p -value = 0.2751).

2. Inheritance state analysis. The availability of nuclear families with two or more offspring enables the identification of inheritance states¹⁰. These represent whether the offspring inherited the same alleles from both parents (“identical” state), the same allele from one parent but different alleles from the other (“haploidentical” state, maternal or paternal according to the parent from which the same allele was inherited), or different alleles from both parents (“nonidentical” state). Some family genotypes are consistent with all inheritance states (e.g. when all family members are homozygous)

and are thus non-informative. Some family genotypes are consistent only with a subset of the inheritance states. For example if both parents are heterozygous A/G, and both offspring are homozygous A/A, clearly the offspring inherited the same alleles from both parents, which is consistent only with the “identical” state. Some family genotypes are consistent with two inheritance states. By combining the evidence from individual SNPs along a chromosome, it is possible to identify contiguous blocks of consistent inheritance, bounded by recombination events in either parent (Figure 3). The overall fraction of the genome present in each inheritance state (Figure 3 inset) deviates little from the expected 25%.

We found that 23andMe genotypes are sufficient for performing this analysis and identifying well-defined inheritance state blocks,

despite covering a very small fraction of the genome (approx. half a million shared SNPs, limited by the lower density version used for Son). This is due to the largely uniform sampling of SNPs along the genome.

3. *Admixture analysis.* Visualization of admixture for Mother and Father was done with ADMIXTURE¹² in the context of other similar Southern European individuals. Figure 2 showed an admixture mapping for a selection of individuals from the Eurogenes Genetic Ancestry project¹³. Mother was denoted as ES7 (red arrow) and Father as ES8 (blue arrow). Mother and Father seemed to be markedly different yet in and around the Portuguese and Spanish cluster of individuals.

Combining SNP chip data and exome SNPs in Son for SNPedia annotation

To leverage all genotype data contained in all different sources, we combined the SNP chip data v2 (574,406 SNPs) with those found in Son's exome data. 10,203 genotypes were annotated in SNPedia when exome SNPs were pooled with the SNP chip version 2 (generated on 11th June, 2012). Processing just the exome, only 925 genotypes were found annotated in SNPedia. It is not surprising that so few additional SNPs were found, as the exome comprises a very small percentage of the total genome. Two SNPs were discovered to have conflicting genotypes between the two platforms: *rs12344615* reported as 'AG' and 'GG' and *rs2290272* reported as 'CT' and 'TT' respectively. The most likely informative SNPs from the exome data are summarized as judged by their observed frequency in HapMap⁸.

SNPs from the Son's exome data

1 Data File

<http://dx.doi.org/10.6084/m9.figshare.92819>

Exome data summary statistics

For the analysis of similarities between genotypes only 23andMe data was analyzed. Although exome sequencing is not widely marketed yet within the DTC providers as an option, this is likely to change in the near future. With our current budget constraints, we were able to sequence Son's exome and relevant SNP and variation data was added for further analysis. A BAM file was created out of 4 FASTQ files downloaded from a server from the Beijing Genomics Institute. A compressed VCF file was also created. A total of 2.54 Gigabases of sequence was aligned at high quality. Summary metrics of the exome were calculated using Picard¹⁴ and showed a minimum of 61.42% of the on-target regions were covered with a depth at least 20x. Genotyping with GATK¹⁵ identified 37,702 variations relative to the reference genome (GRCh37). This was noted to be lower than expected if additional samples had been genotyped concurrently. 97% of the variants identified were within a known gene, 58% of them overlapped a protein domain and 5,565 were non-synonymous (15%) with serious predicted consequences on the protein product (as determined by SIFT¹⁶, PolyPhen¹⁷ or Condel¹⁸). Of these 5,565 potentially pathogenic SNPs, 413 had not been previously identified (verified against dbSNP release 132). This represents a normal figure for the private, novel, non-synonymous

changes carried by most individuals. No more serious, novel changes were identified (such as stop codons gained or lost).

Visualization with SNPedia tools

We used SNPedia's Hilbert curve visualization tool to compare chromosomes between different individuals. Figure 5 shows an example of a Hilbert curve comparison between chromosome 1 of Mother and all family members and the non-CEU individual. Each pixel corresponds to a SNP, colored according to four categories: match (light blue), half-match (dark blue), mismatch (red) and no data (grey). Two patches of light blue are apparent in the Mother-Aunt comparison of chromosome 1, and more appear in other chromosomes (not shown). These correspond to 'identical' segments in which the Mother and the Aunt inherited identical haplotypes from both their parents.

ISCA analysis for the quartet (missing grandfather), (missing grandmother), mother and aunt

1 Data File

<http://dx.doi.org/10.6084/m9.figshare.92820>

The next most similar graph can be seen to be the Mother/Daughter comparison. Much of the graph comparing Mother and Son is grey, representing SNPs present only on 23andMe v3 but absent from v2.

Inference of phenotypes using SNP data

We inferred phenotypes from the observed genotypes, by comparing to all available SNPedia SNP annotations. We analyzed family genotypes using the Promethease tool¹⁹, which allows annotation of observed genotypes from DTC analyses with SNPedia-annotated

23andMe SNPs for which SNPedia annotations are available

5 Data Files

<http://dx.doi.org/10.6084/m9.figshare.92757>

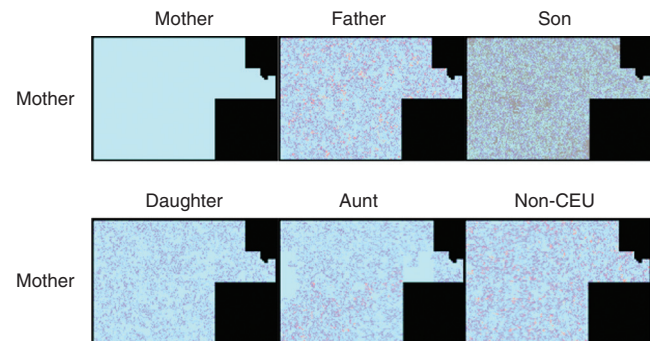


Figure 5. A graphical representation of the SNPs from chromosome 1 for Mother compared with herself, Father, Son, Sister, Aunt and non-CEU. Each pixel represents a SNP. Light blue represents match, dark blue half-match and red conflict. SNPs in Son that are not present in the genotypes of the other individuals are represented in grey.

phenotype associations. These are collected directly from the scientific literature. All family members were analyzed using Promethase and associated phenotypes were collected for further interpretation. SNPedia’s annotation *Magnitude*, denoting a subjective degree of importance for an observed trait as judged by the curator, was used to discriminate between traits that should be considered further in our analyses. A magnitude of 0 denotes

a common genotype for which no associated phenotypic data is known. Magnitude >3 is defined as ‘probably’ interesting. The maximum number is 10. In order to compare family phenotype annotations, all SNPs or ‘genosets’ (groups of SNPs) with equal or greater than magnitude 3 were extracted and summarized in Table 5. Results for Son are not directly comparable as he has fewer SNPs analyzed.

Table 7. A comparison of all SNPedia annotations with Magnitude >= 3 for all family members. Traits have been classified according to the general condition they relate. Red boxes are indicative of a particular phenotype being predicted in the individual. Descriptions for every matched phenotype, extracted directly from SNPedia, are shown in the right column.

Condition	Mother	Father	Daughter	Aunt	Son	Phenotype
Baldness						7x risk of baldness according to an article in Nature. That site may require paid access; the abstract is accessible.
						2x increased risk of baldness 2x increased risk of baldness
Diabetes						Increased risk for type-2 diabetes
						1.3x increased risk for type-2 diabetes
Cardiovascular/ Thrombosis						1.7x increased risk for heart disease. People with this genotype and a long history of high blood sugar are at 7x risk of CAD
						1.5x increased risk for CAD; 1.5x higher risk for coronary artery disease
						7.3x increased risk of hypertension
						Watch out for high fat in diet
Cancer						2.6 times higher odds of developing early stent thrombosis
						Increased risk of various types of cancer. This variant increases risk of numerous types of cancer in many studies. It is in a microRNA
Metabolism						2–3x higher prostate cancer risk if routinely exposed to the pesticide fonofos
						You have 2 variations in MTHFR which influence homocystine levels. People with gs193 are more strongly affected.
						Impaired NSAID drug metabolism, which is a risk factor for gastrointestinal bleeding when taking any of these medications: aceclofenac, celecoxib, diclofenac, ibuprofen, indomethazine, lornoxicam, meloxicam, naproxen, piroxicam, tenoxicam and valdecoxib. You have one of these *CYP2C8*3 (rs11572080 and rs10509681) *CYP2C9*2 (rs1799853) *CYP2C9*3 (rs1057910)
						CYP2C19 Intermediate Metabolizer. Your body breaks down some medicines at a slightly slower than normal rate (which is represented by gs150). Individuals with gs152 genotypes have even slower metabolism. *anti-epileptics (such as diazepam, phenytoin, and phenobarbitone) *anti-depressants (such as amitriptyline and clomipramine) *anti-platelet drug clopidogrel (Plavix) *anti-ulcer proton pump inhibitors like omeprazole (trade names Losec and Prilosec), esomeprazole (trade name Nexium), and lansoprazole (Prevacid) *hormones (estrogen, progesterone).
Warfarin Metabolism						Higher odds of alcoholic liver disease, increased liver fat alcohol seems to be 3x more damaging to your liver than typical. Higher risk for developing fatty liver, fibrosis, and fibrosis progression, with a per allele odds ratio of 2.55, 3.13 and 2.64, respectively. news
						Approximately 30% of people are intermediate metabolizers of the popular anticoagulant Warfarin and would probably need a decreased dosage. This due to rs1799853 or rs1057910 respectively leading to the CYP2C9*2 or CYP2C9*3 alleles. For prodrugs that require activation by CYP2C9, an alternative treatment or increased dose should be considered. See also gs126
						Probably impaired Warfarin metabolism.
Miscellaneous						Approximately 7–10% of people are poor metabolizers of the popular anticoagulant Warfarin and would probably need a decreased dosage. This due to mutations in rs1799853 or rs1057910 causing an inactive CYP2C9 gene. You are at increased risk of drug-induced side effects due to diminished drug elimination. Prodrugs dependent on CYP2C9 metabolism may fail to generate the active form of the drug.
						Substantially increased odds of developing V617F-positive MPN.
Miscellaneous						You are heterozygous at all 3 of the SNPs which are known to influence the ability to taste bitterness. This means you are better than average at detecting bitter tastes while young, but that this ability will decrease to less than average during adulthood. As a child you will probably hate brussel sprouts, and by early adulthood will discover that olives and brussel sprouts now taste good. A 2010 study shows the change bitter sensitivity which occurs over the lifespan (from bitter sensitive to less so) is more common in people with this genoset. Children with this genotype could perceive a bitter taste at lower PROP concentrations than could heterozygous adults. The threshold for adolescents was intermediate. The 3 SNPs are rs10246939, rs1726866, rs1713598 in the gene TAS2R38.

Based on observed results, the maternal hereditary line seems to carry greater risks related to diabetes and cardiovascular/thrombosis related conditions. In addition, both males (Son and Father) have greater risk of baldness as well as mixed results in terms of their ability to metabolize drugs. All the family shares a common trait of substantially increased odds of developing V617F-positive Myeloproliferative neoplasms.

Discussion

We have presented here, to our knowledge, the first systematic analysis of DTC genomic data using non-commercial or low cost resources, combining data from 5 blood-related family members. The purpose of this study does not lie in uncovering the phenotypic predictions or genetic findings in these individuals' genomes. Instead we aimed to demonstrate to what extent, in principle, any individual can interpret their personal genome using only public resources with an affordable budget and no laboratory equipment. We stress that our goal is not to diminish the value of DTC industry services, which have catalyzed the access to, and interest in, personal genomics data in the wider public. However, as the adoption of DTC personal genomics tests becomes ever more widespread, we envisage community phenotype association and third party public tools to become more significant in the overall interpretation of personal genomics results.

Although we found the no-call rate to be comparable between both versions of the 23andMe platform, Mother had a greater undercall rate than the other individuals. Also, results given by no-call rates suggest that there might be some intrinsic differences between the two chips. Nevertheless the current number of samples analyzed is not large enough to make this conclusion, as only five chips of data analyzed do not provide any basis for their overall performance. Therefore our error estimation rates presented here should only be considered in the context of this analysis and not as representative of the DTC company's overall quality scores.

Based on the no-call rate and the relative ratios of homozygous and heterozygous sites reported, we computed an 'undercall' rate and used it to estimate the number of heterozygous sites mistakenly reported as homozygous. The frequency of such events is small (~0.2%). Nevertheless, the fact that up to 2,000 heterozygous variants may be missed is a reminder that interpretation of personal disease risks should be done with caution. Findings of potential medical relevance should always be verified for correctness. Whenever possible, it is beneficial to perform the analysis in the context of families: identification of MIEs and State Consistency Errors¹⁰ is a powerful tool to assess genotyping quality.

The similar level at which identical genotypes is shared between the relatives and the non-CEU individual is consistent with the ethnically close background for the family members. This, however, does not suggest that Mother or Father are directly related. In fact, when performing an admixture analysis with unrelated individuals of Southern European descent (Figure 2) it is clear that while the parents cluster within reasonably close distance to other Spanish individuals, they display a typical level of genotype sharing between two people from the Iberian Peninsula.

We also indicated that when comparing genotype similarities between siblings and parents, we found that Daughter was significantly closer to Father than Son was. Although the expectation was that

both Daughter and Son should be equally similar to both parents, these unexpected results may be the reflection of bias in the subset of markers used in the DTC analysis. These results are not therefore indicative of Daughter's genome being closer to either parent, as most of the genome is missing and hence any inference in this respect cannot be made. In the context of SNP analysis, however, it is worthwhile reporting such SNP differences as these will influence the overall results reported back to the DTC customer. This in turn may explain why observed susceptibility risks vary among family members when comparing their phenotypic annotations.

The identification of blocks of identical, haploidentical or nonidentical genotype between family members (e.g. Son and Daughter in Figure 4), highlights the location of meiotic recombinations. These blocks provide well defined expectations for whether the genes included in them should display similar or distinct genotypes. This information is valuable in the context of genetic research¹⁰ but also to the general public, to predict shared phenotypes among family members. Within 'identical' blocks, siblings are essentially identical twins: this fact gains special personal meaning in the context of DTC genetic analysis.

Publicly curated data, like that available in SNPedia, is exposed to error due to human mistakes or malicious intent. Although this is a legitimate concern, it has been shown that with similar public annotation resources (Wikipedia for example), vandalism rates are very low in areas of specialist academic interest²⁰. As these open access resources mature and grow, maintaining accuracy will be an important consideration for the contributing community.

In this study we have not found any annotated genotype that is likely to raise significant health concerns among the family individuals. It is inevitable, however, that as more individuals investigate their own personal genomes, and more statistical associations are uncovered, genotype/phenotype correlations with serious health implications will be accessible through open access resources. Interpreting such information appropriately is bound to be difficult for individuals who are not expert geneticists; this poses special ethical challenges from the point of view of how to present the data. DTC companies have recognized this ethical issue by implementing additional access controls to some particularly sensitive annotations, such as those genotypes associated with a high risk of developing Alzheimer's disease or some forms of breast cancer. To our knowledge, public annotation resources do not currently make such distinctions. We therefore urge caution when investigating personal genomic data for health risks, especially when using open access information. We recommend those who uncover personal genomic information of medical concern seek the advice of genetic counselors, who can interpret and advise on the context of what it really means to the tested individual.

Methods

SNP chip data

Five 23andMe genome analysis kits were purchased at two time points. Son's kit was bought in May 2009 and was tested with 23andMe version 2 (~576,000 SNPs). The other 4 family members were analyzed in one batch, with kits bought in December 2010 and results returned in February 2011. The second batch used 23andMe's chip version 3 with ~967,000 SNPs per genome analyzed. After discussion of results, consent was given by all family members of the family to publish their genotypes.

Calculation of error rates

Assume N SNPs were tested, and of these, fN are truly heterozygous. We wished to compute f (heterozygous fraction) from the observed numbers of homozygous, heterozygous and failed SNPs.

For each SNP with a dbSNP 'rs' identifier, we assumed that 1) the SNP is biallelic, 2) it is present in diploid state. We further assumed a probability x of not observing a given allele (the "undercall rate"), and assumed this probability was equal for both alleles, at all sites. Finally, we assumed that the probability of observing a wrong allele was zero ("overcall rate").

If the true state of the SNP was heterozygous, the following could have happened. 1) If neither allele was observed (double undercall), the SNP was called "NULL" (with conditional probability = x^2). 2) If one allele was not observed (single undercall), the SNP was called "HOM" (conditional probability = $2x$). 3) If both alleles were observed, the SNP was called "HET" (conditional probability = $1-2x-x^2$).

If the true state of the SNP was homozygous, there was only one type of allele to be observed. Thus, the following could have happened. 1) If the allele was not observed, the SNP was called "NULL" (conditional probability = x). Otherwise, 2) the SNP was called "HOM" (conditional probability = $1-x$).

The expected frequencies for NULL, HET and HOM were:

- NULL' = $fx^2+(1-f)x$
- HET' = $f(1-2x-x^2)$
- OM' = $(1-f)(1-x)+2fx = 1-x+f(3x-1)$

Since HOM + HET + NULL = 1, the undercall rate x was given by solving:

$$x^3 + (1+\text{HOM})x^2 + (3\text{HET}+2\text{HOM}-3)x + \text{NULL} = 0$$

The heterozygous fraction f was then given by: $f = \text{HET}/(1-2x-x^2)$, and the number of "missing" heterozygous sites could be computed by: missing = $(f-\text{HET})N$. Finally, the number of heterozygous sites reported as homozygous was given by:

$$\text{het2hom} = \text{missing} - N*\text{NULL}* \text{HET}/(\text{HET}+\text{HOM})$$

Extraction of SNPs from exome data

Raw reads were aligned to the reference GRCh37 using bwa 0.61²¹. Local realignment was performed around indels with the Genome Analysis Toolkit (GATK v1.4)¹⁵ framework for variation discovery and genotyping using next-generation DNA sequencing data. Optical and PCR duplicates were marked in BAM files using Picard 1.62¹⁴. Original HiSeq base quality scores were recalibrated using GATK TableRecalibration and variants called with GATK UnifiedGenotyper. Indels and SNPs were hard-filtered according to Broad Institute best-practice guidelines²² to eliminate false positive calls and produce the final VCF.

Son exome files

7 Data Files

<http://dx.doi.org/10.6084/m9.figshare.92584>

Phenotype inference using promethease

Promethease, a SNPedia tool for phenotype inference, was used for assignment of SNPedia annotations to observed SNPs. SNPedia annotations contain manually curated SNPs that summarize phenotype associations observed in a particular population. Phenotype associations were inferred using SNPedia's SNP ids, which correspond to dbSNP²³. In our analysis only SNPedia annotations of ≥ 3 Magnitude were examined. Magnitude is a subjective score that helps prioritize SNPs according to their expected importance and the phenotypic annotation itself in the form of free text. Magnitude is assigned by SNPedia entry curators.

Calculation of similarity scores

We calculated similarity scores using our own Perl scripts and MySQL. Similarity between any given two individuals was calculated as the total number of matching SNPs plus half-matches divided by 2. Negative results were counted as the total number of conflicts plus the number of half-matches divided by 2. As the platforms used are different for Son as compared to the rest, Son can only be compared relative to himself and not against other individuals. Similarities between individuals were statistically tested using the R package for Pearson's Chi-squared with Yates' correction.

Consent

Written consent for publication of their genotype and phenotype was obtained from all family individuals.

Author contributions

Manuel Corpas and Darren Logan conceived the experiments. Manuel Corpas, Gustavo Glusman and Darren Logan wrote the paper with input from the other authors. Gustavo Glusman, Mike Carriaso, Manuel Corpas, Daniel Swan contributed with data and analyses, Rafael Jimenez, Bastian Greshake and Jong Bhak contributed comments/suggestions and their expertise of the field. All authors read and approved the final version.

Competing interests

Michael Carriaso is the developer of Promethease and SNPedia. Some aspects of the Promethease analysis shown here cost \$2 per individual.

Acknowledgements

We are grateful to the Eurogenes Genetic Ancestry Project for providing the admixture analysis of individuals from (Southern Europe [Figure 2](#)).

Grant information

The author(s) declared that no grants were involved in supporting this work.

References

1. Saxena R, Voight BF, Lyssenko V, *et al.*: **Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.** *Science.* 2007; **316**(5829): 1331–1336.
[PubMed Abstract](#) | [Publisher Full Text](#)
2. Phillips C, Salas A, Sánchez JJ, *et al.*: **Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs.** *Forensic Sci Int Genet.* 2007; **1**(3–4): 273–280.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Carriaso M, Lennon G: **SNPedia: a wiki supporting personal genome annotation, interpretation and analysis.** *Nucleic Acids Res.* 2012; **40**(Database issue): D1308–D1312.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
<http://opensnp.org/>
4. <http://www.mendeley.com/>
5. <http://www.mendeley.com/>
6. Huss JW III, Lindenbaum P, Michael M, *et al.*: **The Gene Wiki: community intelligence applied to human gene annotation.** *Nucleic Acids Res.* 2010; **38**(Database issue): D633–D639.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Good BM, Clarke EL, Loguercio S, *et al.*: **Linking genes to diseases with a SNPedia-Gene Wiki mashup.** *J Biomed Semantics.* 2012; **3**(Suppl 1): S6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Gibbs RA, Belmont JW, *et al.*: **The international HapMap project.** *Nature.* 2003; **426**(6968): 789–796.
[PubMed Abstract](#) | [Publisher Full Text](#)
<http://www.23andme.com/>
9. <http://www.23andme.com/>
10. Roach JC, Glusman G, Smit AF, *et al.*: **Analysis of genetic inheritance in a family quartet by whole-genome sequencing.** *Science.* 2010; **328**(5978): 636–639.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. <http://customer.care.23andme.com/entries/21259007-what-does-no-call-not-genotyped-mean-in-browse-raw-data>
12. Alexander DH, Novembre J, Kennedy L, *et al.*: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome Res.* 2009; **19**(9): 1655–1664.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
<http://bga101.blogspot.com/>
13. <http://picard.sourceforge.net/>
14. <http://picard.sourceforge.net/>
15. McKenna A, Hanna M, Eric B, *et al.*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res.* 2010; **20**(9): 1297–1303.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res.* 2003; **31**(13): 3812–3814.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Adzhubei IA, Schmidt S, Leonid P, *et al.*: **A method and server for predicting damaging missense mutations.** *Nat Methods.* 2010; **7**(4): 248–249.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. Gonzalez-Perez A, Lopez-Bigas N: **Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel.** *The American Journal of Human Genetics.* 2011; **88**(4): 440–449.
[Publisher Full Text](#)
19. <http://www.snpedia.com/index.php/Promethease>
20. Finn RD, Gardner PP, Bateman A, *et al.*: **Making your database available through Wikipedia: the pros and cons.** *Nucleic Acids Res.* 2012; **40**(Database issue): D9–D12.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
21. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2009; **25**(14): 1754–1760.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v3
23. Sherry S, Ward MH, Kholodov M, *et al.*: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res.* 2001; **29**(1): 308–311.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 26 July 2012

<https://doi.org/10.5256/f1000research.103.r201>

© 2012 Gilissen C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Christian Gilissen

Department of Human Genetics, Radboud University Nijmegen, Nijmegen, The Netherlands

This is a well written and interesting article that describes analyses on genome-wide genotyping information from a DTC company.

Major concerns

- Although the analyses are carried out very well, the overall purpose and conclusion of the paper is rather mixed and somewhat unclear. The authors state that they “explore the extent to which phenotype inference and genotype analysis can be carried out solely using existing public or very low cost resources.” or as in the abstract “providing an evaluation of the current potential of public available analysis tools for personal genomics”.
- The discussion section focuses very much on the results of the different analyses, whereas that is (according to the purpose of the paper) not that relevant. Rather in that context it would be more interesting to speculate on other aspects (for example, how easy it is for a typical DTC customer to apply the publicly available tools, or how easy it is to (correctly) interpret the outcome of such tools) and to provide a real conclusion on “the extent to which phenotype inference and genotype analysis can be carried out solely using existing public or very low cost resources.”
- It is not clear whether the tools that were investigated compromise all public domain tools for these kinds of analyses. If the authors do not investigate all available tools, the authors should motivate their choice for the described tools

Given the purpose of the paper, a table with the available tools and their “potential” would be very useful.

Minor concerns

- The fact that “Son” was done with a V2 assay complicates all comparisons in the paper. It seems to paper would be a lot more straightforward if the authors used a V3 assay for the son as well.
- Page 7, “Exome data summary statistics”, authors mention that the numbers they find are what is to be expected. Please add a reference to substantiate this.
- Page 10, the cond. prob. of a heterozygous call for a heterozygous SNP reads “ $1-2x-x^2$ ” and should probably be “ $1-2x-x^2$ ”. Similar with the expected frequency for NULL’.

Summary

- Is the title appropriate for the content of the article? **Yes**
- Is the abstract a suitable summary of the article? **Yes**
- Is the article well constructed and clear? **Yes, the article is well constructed, but could do with a conclusion and discussion that is more in line with the purpose of the paper. (see comments above)**
- Is there adequate analysis, including information on how the data were analyzed (e.g. programs, code, stats etc.)? **Yes**
- Are the conclusions sensible and balanced? **The conclusions does not answer the questions that were asked at the outset of the paper.**
- Have any potential biases or competing interests been disclosed? **Yes**

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 23 July 2012

<https://doi.org/10.5256/f1000research.103.r200>

© 2012 Robinson P. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Peter N. Robinson

The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA

23andMe and several other companies have been offering Direct-to-consumer (DTC) genomic testing for several years now. In general, such tests seem to have a bad reputation amongst human geneticists because of the perception that individuals may not be able to interpret the findings in a useful way.

Additionally, although genetic testing has an important and widely accepted role in the diagnostics of Mendelian disease and several other areas, many feel that the clinical utility of genomic testing for common disorders such as heart disease or hypertension has not been shown to date. For instance, the Evaluation of Genomic Applications in Practice and Prevention Working Group (EWG) found insufficient evidence to recommend testing for the 9p21 genetic variant or 57 other variants in 28 genes to assess risk for cardiovascular disease in the general population ([Genetics in Medicine \(2010\) 12, 839–843](#)).

Be that as it may, has become well known and the range of variants now tested by companies such as 23andMe is increasing; with the advent of relatively cheap exome and even genome sequencing, it seems quite likely that DTC exome and genome sequencing will be offered in the not too distant future. Therefore, it is important to understand several aspects about DTC testing, including the quality of the raw product (i.e., are the genotype calls correct), the depth, correctness, and utility of annotations provided by DTC companies or widely available to the public, and perhaps most importantly, to know how typical consumers of DTC products use the findings for their own health care or, say, for planning life-style modifications.

This paper offers a detailed look at several of these aspects, and I think does an excellent job at providing the reader with a sense of what kind of data consumers of DTC products can expect. I think the paper would profit from a number of minor revisions. The paper essentially deals with two different topics. The first involves a number of computational quality control procedures that require a good deal of bioinformatics expertise (I doubt that much of this would be in the reach of most of the customers of DTC sequencing), ranging from a calculation of error rates, an analysis of Mendelian inheritance errors, genotype distributions, and admixture analysis. It was not entirely clear to me how the p-values reported for the SNP similarity analysis were calculated, and a more precise definition of how pseudo-MIEs were counted should be provided. The authors should provide full methodological details of how this analysis was performed. It is also something of a distraction that one of the samples was analyzed using the 23andMe v2 kit, while the rest of the family was analyzed using the v3 kit. I imagine simply that the Son was the person who “went first”, but the apparently different error characteristics of the two versions limit somewhat the findings of the paper.

The second major aspect of this paper involves analysis that could be done by non-specialists, using tools such as SNPedia. It was not entirely clear to me why the authors performed the analysis in Figure 5. The heading called “Inference of Phenotypes Using SNP data” is misleading because actually what the findings reveal is a genotype, say, a risk for going bald. However, if the son is not currently bald, then he does not now have the phenotype “baldness”. Instead, the only thing that can be inferred from the genotype is an increased risk of baldness. This is a very important distinction, and it is very important especially for the general public to realize that having a genotype that is associated with an increased risk of some phenotype does not necessarily mean that one will actually develop that phenotype.

It would be a nice addition to this paper to hear more about how the family reacted to these findings. Have the increased risks for some of the diseases mentioned led to life style changes? Did the family members report the findings to their physicians? Was a family member worried or uncomfortable about hearing the findings?

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research