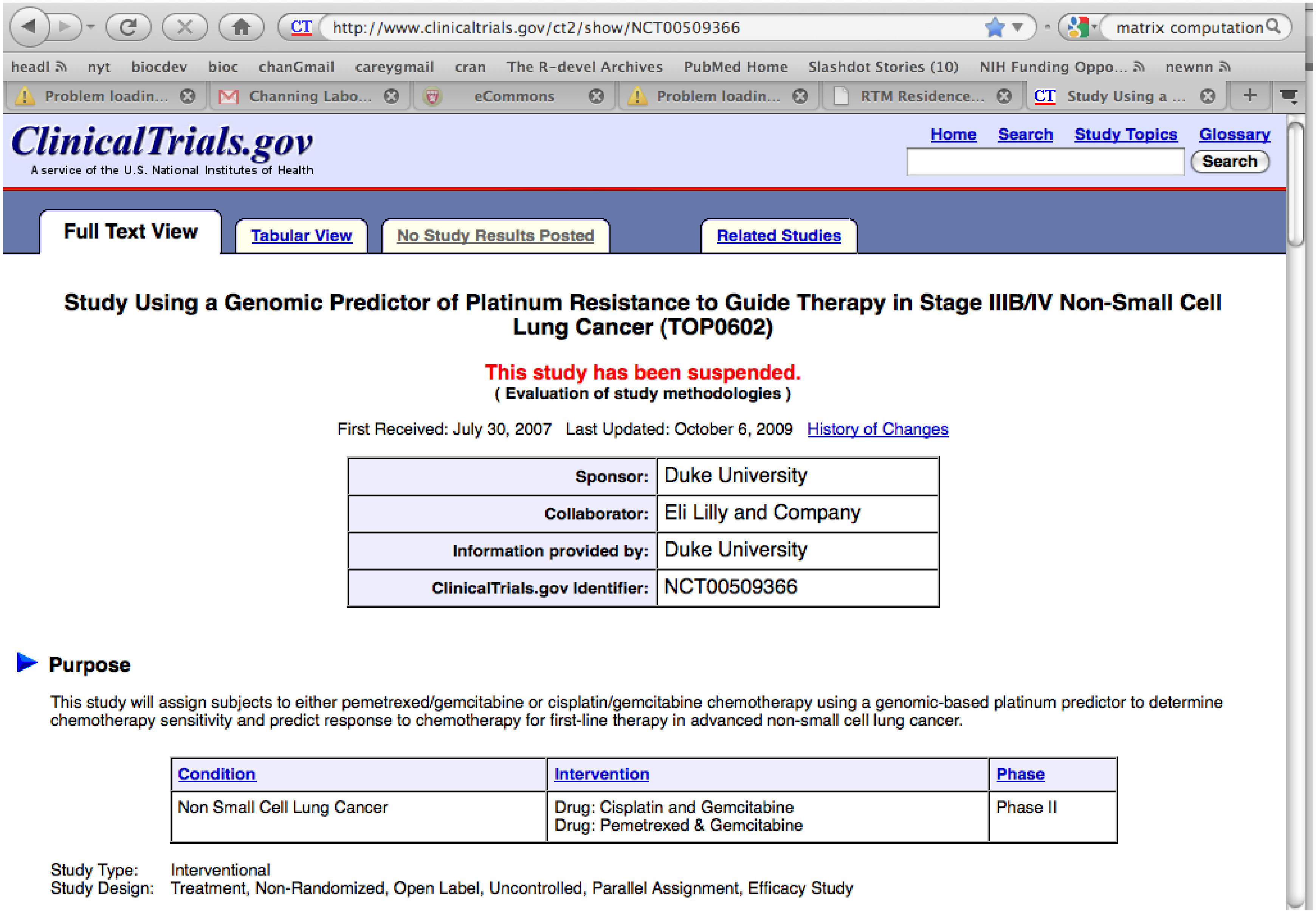
Bioconductor approaches to NGS: Rare variants

VJ Carey, Channing Laboratory, Brigham and Women's Hospital

- some background
  - epidemiologic concepts
  - technical issues in identifying variants with NGS
- annotation and filtering resources
- exploratory analysis of rare variant existence and impact

Take home message

- Don't believe in magic.
- Nontrivial computational work always contains errors
- ... until it has been tested, verified
- ... even then errors may persist
- testing discipline, sanity checks, vigilance of users – we need more
- (Polyglots: What is the translation of "slog" into your favorite language?)



ClinicalTrials.gov
A service of the U.S. National Institutes of Health

Full Text View | Tabular View | No Study Results Posted | Related Studies

Study Using a Genomic Predictor of Platinum Resistance to Guide Therapy in Stage IIIB/IV Non-Small Cell Lung Cancer (TOP0602)

This study has been suspended.
(Evaluation of study methodologies)

First Received: July 30, 2007 Last Updated: October 6, 2009 History of Changes

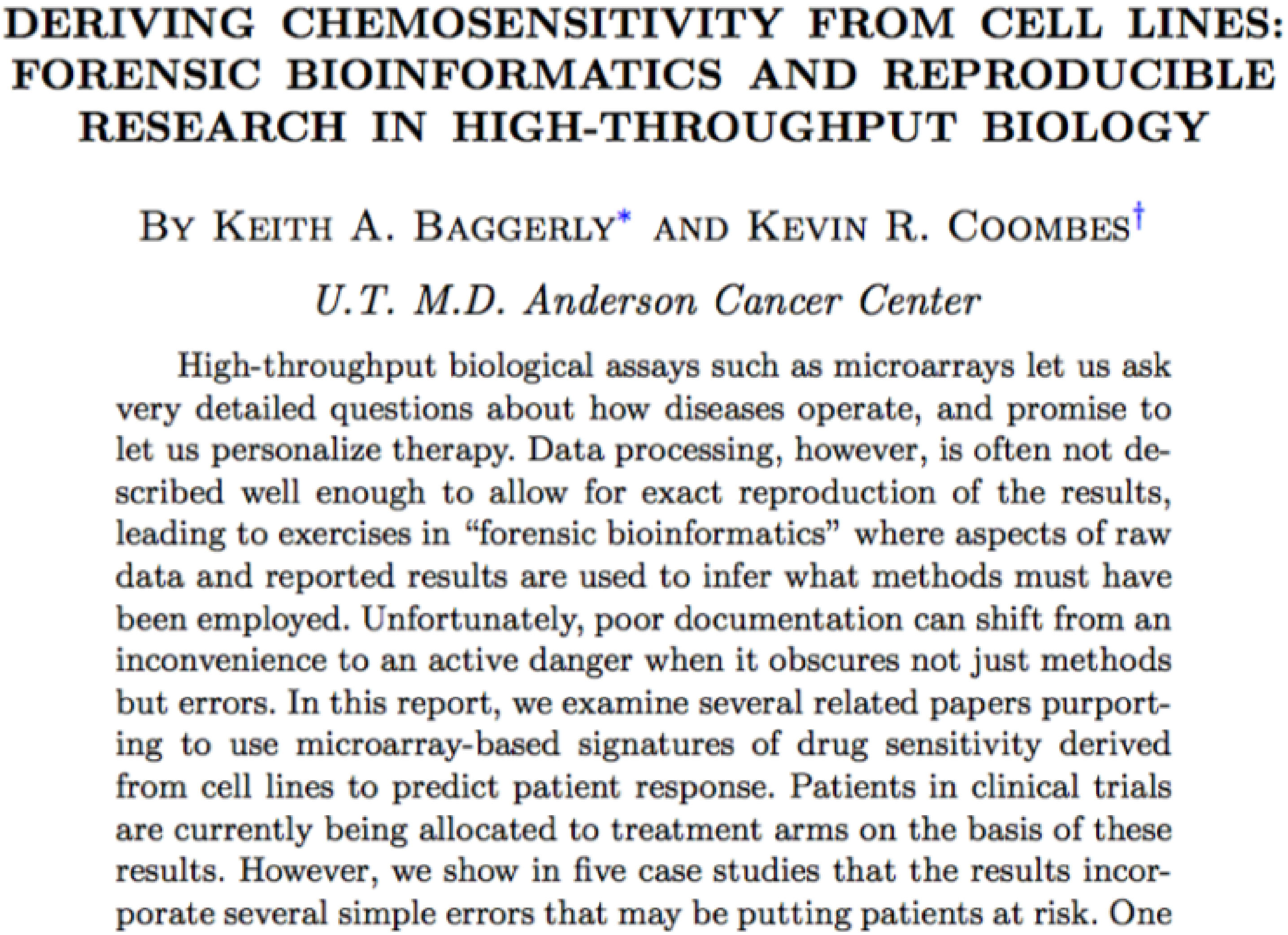
Sponsor:	Duke University
Collaborator:	Eli Lilly and Company
Information provided by:	Duke University
ClinicalTrials.gov Identifier:	NCT00509366

Purpose

This study will assign subjects to either pemetrexed/gemcitabine or cisplatin/gemcitabine chemotherapy using a genomic-based platinum predictor to determine chemotherapy sensitivity and predict response to chemotherapy for first-line therapy in advanced non-small cell lung cancer.

Condition	Intervention	Phase
Non Small Cell Lung Cancer	Drug: Cisplatin and Gemcitabine Drug: Pemetrexed & Gemcitabine	Phase II

Study Type: Interventional
Study Design: Treatment, Non-Randomized, Open Label, Uncontrolled, Parallel Assignment, Efficacy Study



DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

By Keith A. Baggerly* and Kevin R. Coombes†

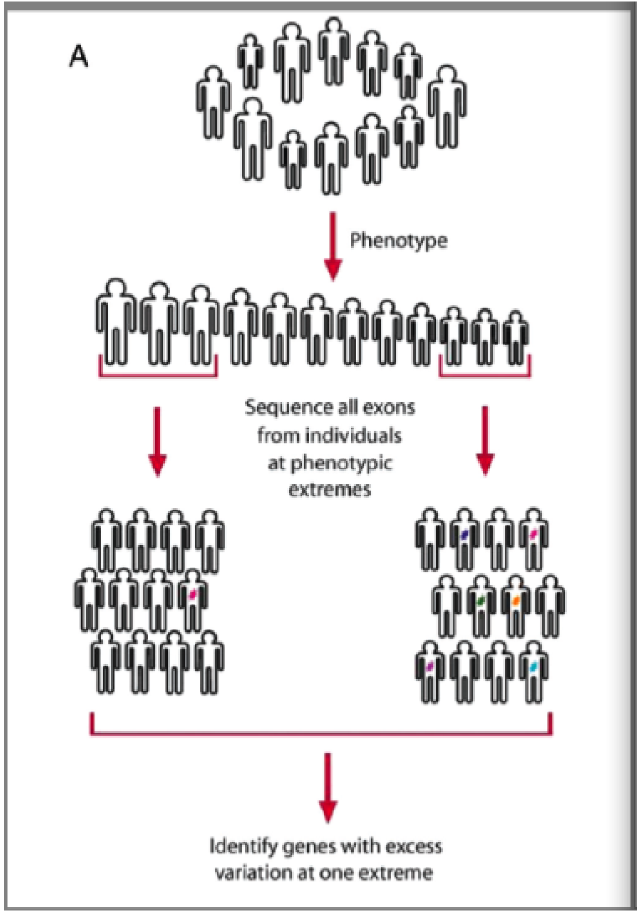
U.T. M.D. Anderson Cancer Center

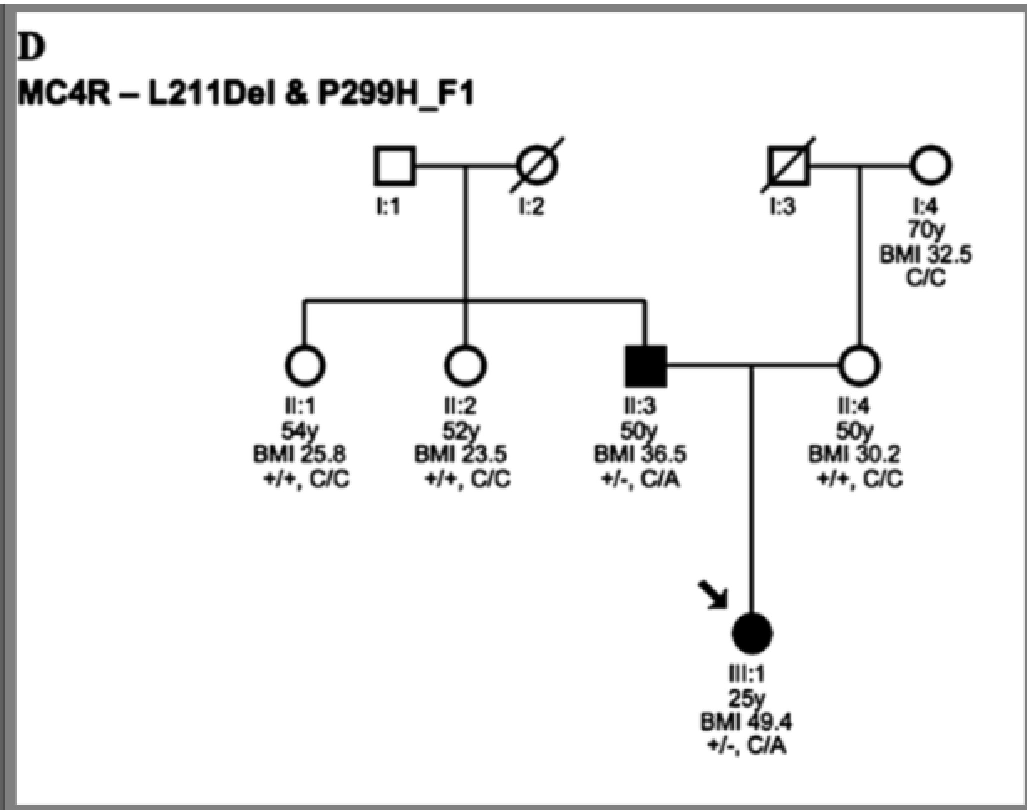
High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic bioinformatics" where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report, we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One

Conclusion on reproducibility

- Open-source, platform-independent computing tools can solve serious problems
- This course attempts to build your versatility in confronting very complex problems of interpretation
- Code, data and metadata can be in error
- When Microsoft/Apple/Linux discovers a bug, they go out onto your computer and try to fix it (windows/macosx/synaptic updates)
- We certainly can't do that; and in science, reconstruction of flawed analyses can be very hard
- Versioned packages of data and code can help manage complex analytic chains

Kryukov+, PNAS 2009; Ahituv+ AJHG 2007





Published in final edited form as:
Nat Genet. 2008 June ; 40(6): 695–701. doi:10.1038/ng.f.136.

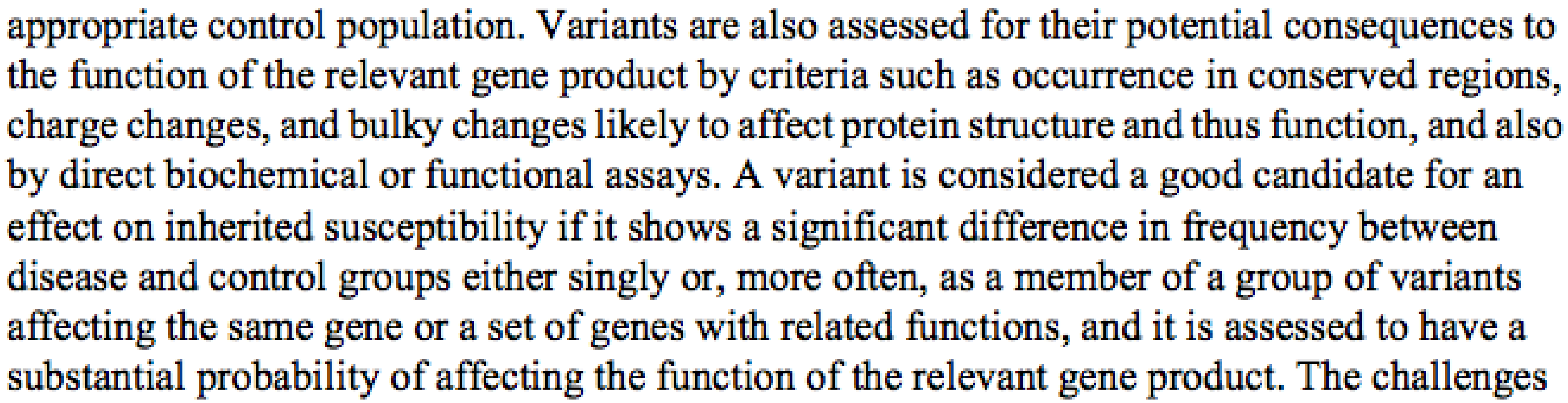
Common and rare variants in multifactorial susceptibility to common diseases

Walter Bodmer and **Carolina Bonilla**
Walter Bodmer and Carolina Bonilla are at the Cancer and Immunogenetics Laboratory, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK.

The rare variant hypothesis: colorectal cancer as a model

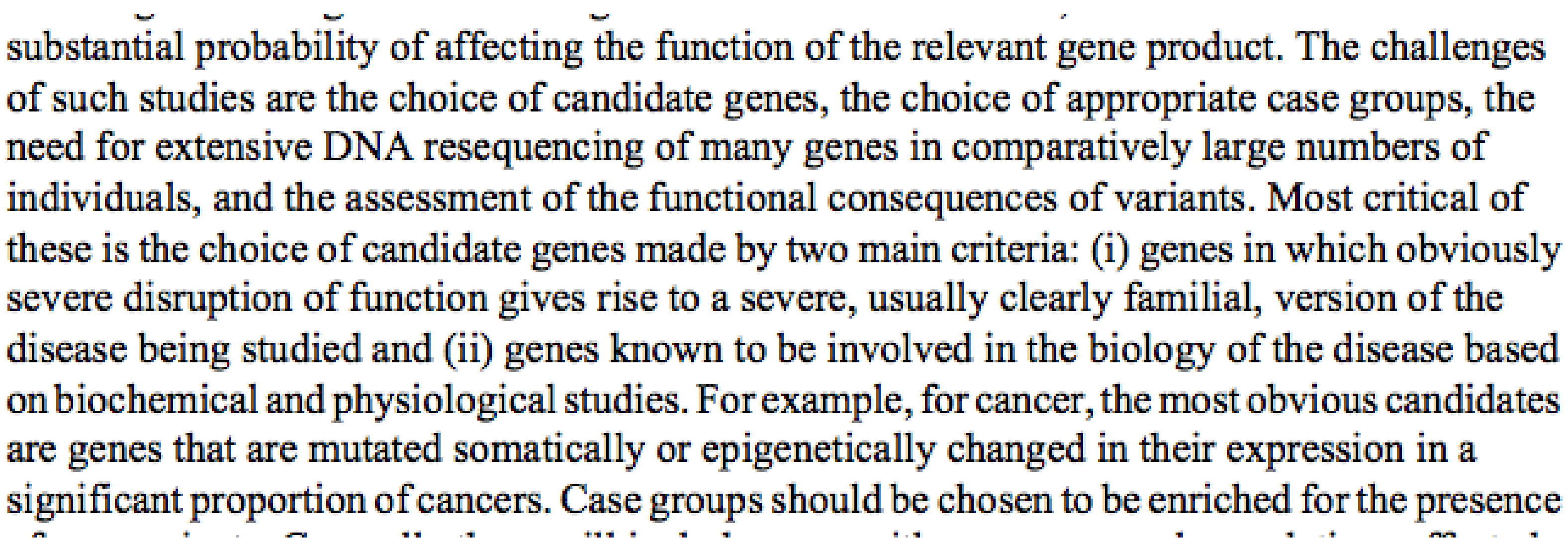
About 5% of cases of colorectal cancer (CRC) are associated with inherited, dominant, familial mendelian susceptibility, especially FAP (familial adenomatous polyposis), caused by severely deleterious highly penetrant mutations in the *APC* gene, and HNPCC (hereditary nonpolyposis colorectal cancer), caused by mutations in mismatch repair genes (see ref. 18 for an example). Another 20–30% of cases are thought to be due to inherited susceptibility that is 'multifactorial', namely, associated with much lower penetrance variants that do not give rise to clear-cut familial patterns of inheritance. An important role for rare variants in inherited multifactorial susceptibility to colorectal cancer was first suggested by the effects of rare missense variants in *APC*[19,20]. The biggest gap in our knowledge of the inherited susceptibility to colorectal cancer—as also for essentially all the relatively common chronic diseases—concerns the 20–30% of cases that are multifactorial. It is that gap which WGAS and rare variant studies aim to fill.

Criteria for variants impacting susceptibility



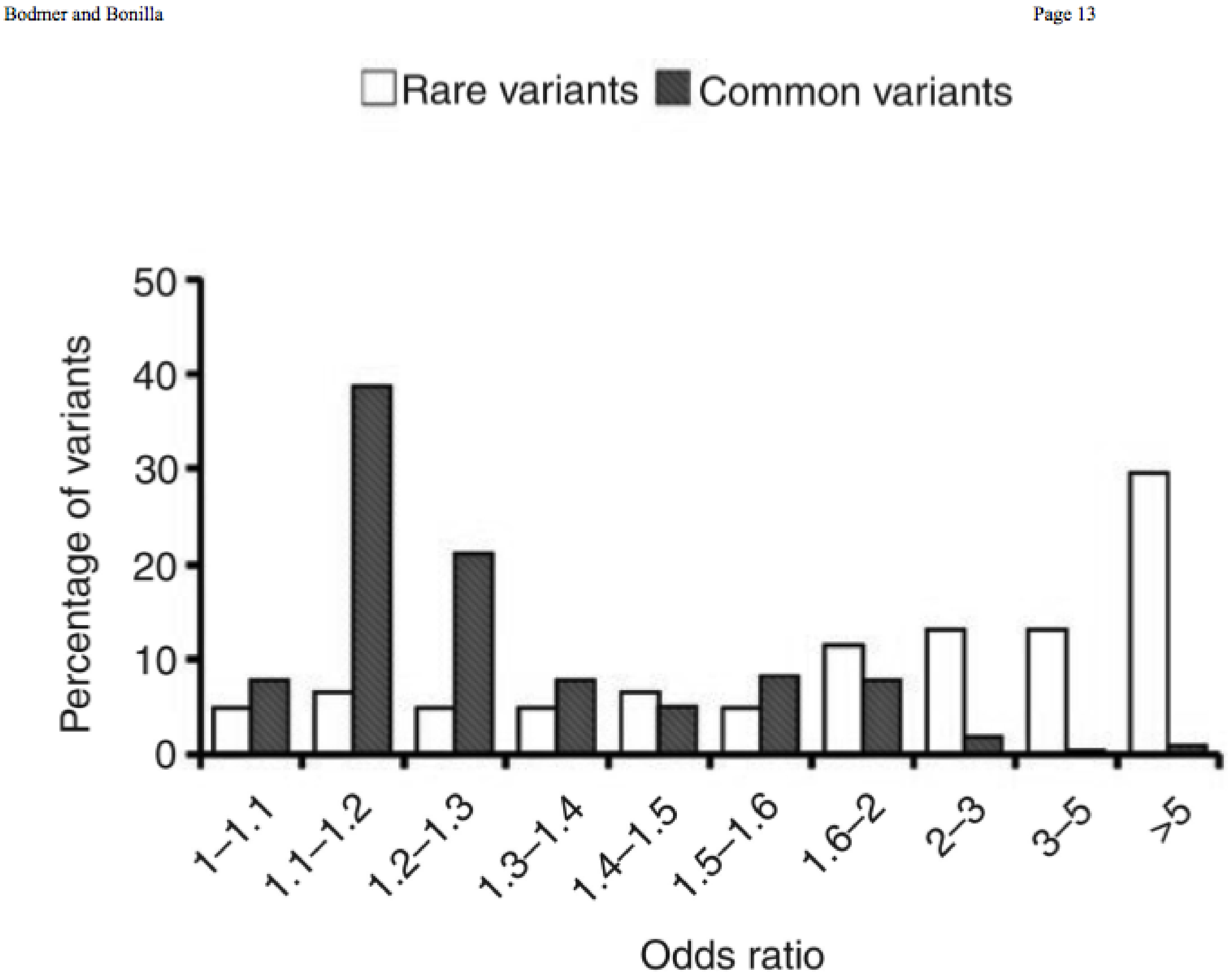
appropriate control population. Variants are also assessed for their potential consequences to the function of the relevant gene product by criteria such as occurrence in conserved regions, charge changes, and bulky changes likely to affect protein structure and thus function, and also by direct biochemical or functional assays. A variant is considered a good candidate for an effect on inherited susceptibility if it shows a significant difference in frequency between disease and control groups either singly or, more often, as a member of a group of variants affecting the same gene or a set of genes with related functions, and it is assessed to have a substantial probability of affecting the function of the relevant gene product. The challenges

Criteria for gene selection for variant searches



substantial probability of affecting the function of the relevant gene product. The challenges of such studies are the choice of candidate genes, the choice of appropriate case groups, the need for extensive DNA resequencing of many genes in comparatively large numbers of individuals, and the assessment of the functional consequences of variants. Most critical of these is the choice of candidate genes made by two main criteria: (i) genes in which obviously severe disruption of function gives rise to a severe, usually clearly familial, version of the disease being studied and (ii) genes known to be involved in the biology of the disease based on biochemical and physiological studies. For example, for cancer, the most obvious candidates are genes that are mutated somatically or epigenetically changed in their expression in a significant proportion of cancers. Case groups should be chosen to be enriched for the presence

A survey of odds ratios estimated as rare/common variant effects



Bodmer and Bonilla Page 13

Summary

- (configurations of) rare variants are receiving increasing attention as vehicles for reasoning about disease etiology and treatment
- development of neutral and disease-enriched variant catalogues is proceeding rapidly
- harvesting and interpreting new high-resolution information on variants is highly technical and not easy to make transparent

Recent results with exome sequencing

nature Vol 461|10 September 2009|doi:10.1038/nature0825

LETTERS

Targeted capture and massively parallel sequencing of 12 human exomes

Sarah B. Ng[1], Emily H. Turner[1], Peggy D. Robertson[1], Steven D. Flygare[1], Abigail W. Bigham[2], Choli Lee[1], Tristan Shaffer[1], Michelle Wong[1], Arindam Bhattacharjee[4], Evan E. Eichler[1,3], Michael Bamshad[2], Deborah A. Nickerson[1] & Jay Shendure[1]

- proof of concept with Freeman-Sheldon syndrome, a rare dominantly inherited disease involving malformation and joint contracture
- employs various sequences from 1000 genomes as neutral reference
- exome sequencing cited as 20-fold less costly than whole-genome
- lose access to roles of noncoding variants

mutations in *MYH3* (ref. 5). Unpaired, 76 base-pair (bp) reads[12] from post-enrichment shotgun libraries were aligned to the reference genome[13]. On average, 6.4 gigabases (Gb) of mappable sequence was generated per individual (20-fold less than whole genome sequencing with the same platform[12]), and 49% of reads mapped to targets (Supplementary Table 1). After removing duplicate reads that represent potential polymerase chain reaction artefacts[14], the average fold-coverage of each exome was 51× (Supplementary Fig. 1). On average per exome, 99.7% of targeted bases were covered at least once, and 96.3% (25.6 Mb) were covered sufficiently for variant calling (≥8× coverage and Phred-like[15] consensus quality ≥30). This corresponded to 78% of genes having >95% of their coding bases called (Supplementary Fig. 2 and Supplementary Data 2). The average pairwise correlation coefficient between individuals for gene-by-gene coverage was 0.87, consistent with systematic bias in coverage between individual exomes.

“High” concordance with illu 1M SNP chip

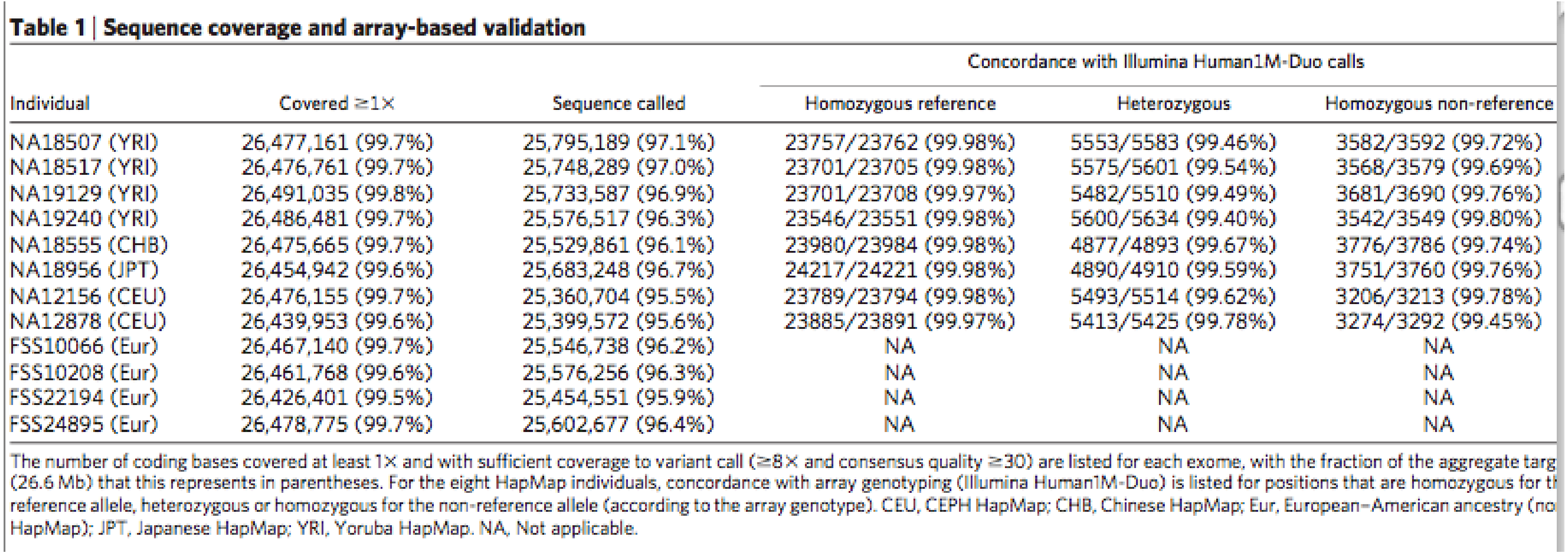


Table 1 | Sequence coverage and array-based validation

			Concordance with Illumina Human1M-Duo calls		
Individual	Covered ≥1×	Sequence called	Homozygous reference	Heterozygous	Homozygous non-reference
NA18507 (YRI)	26,477,161 (99.7%)	25,795,189 (97.1%)	23757/23762 (99.98%)	5553/5583 (99.46%)	3582/3592 (99.72%)
NA18517 (YRI)	26,476,761 (99.7%)	25,748,289 (97.0%)	23701/23705 (99.98%)	5575/5601 (99.54%)	3568/3579 (99.69%)
NA19129 (YRI)	26,491,035 (99.8%)	25,733,587 (96.9%)	23701/23708 (99.97%)	5482/5510 (99.49%)	3681/3690 (99.76%)
NA19240 (YRI)	26,486,481 (99.7%)	25,576,517 (96.3%)	23546/23551 (99.98%)	5600/5634 (99.40%)	3542/3549 (99.80%)
NA18555 (CHB)	26,475,665 (99.7%)	25,529,861 (96.1%)	23980/23984 (99.98%)	4877/4893 (99.67%)	3776/3786 (99.74%)
NA18956 (JPT)	26,454,942 (99.6%)	25,683,248 (96.7%)	24217/24221 (99.98%)	4890/4910 (99.59%)	3751/3760 (99.76%)
NA12156 (CEU)	26,476,155 (99.7%)	25,360,704 (95.5%)	23789/23794 (99.98%)	5493/5514 (99.62%)	3206/3213 (99.78%)
NA12878 (CEU)	26,439,953 (99.6%)	25,399,572 (95.6%)	23885/23891 (99.97%)	5413/5425 (99.78%)	3274/3292 (99.45%)
FSS10066 (Eur)	26,467,140 (99.7%)	25,546,738 (96.2%)	NA	NA	NA
FSS10208 (Eur)	26,461,768 (99.6%)	25,576,256 (96.3%)	NA	NA	NA
FSS22194 (Eur)	26,426,401 (99.5%)	25,454,551 (95.9%)	NA	NA	NA
FSS24895 (Eur)	26,478,775 (99.7%)	25,602,677 (96.4%)	NA	NA	NA

The number of coding bases covered at least 1× and with sufficient coverage to variant call (≥8× and consensus quality ≥30) are listed for each exome, with the fraction of the aggregate targ (26.6 Mb) that this represents in parentheses. For the eight HapMap individuals, concordance with array genotyping (Illumina Human1M-Duo) is listed for positions that are homozygous for tl reference allele, heterozygous or homozygous for the non-reference allele (according to the array genotype). CEU, CEPH HapMap; CHB, Chinese HapMap; Eur, European–American ancestry (no HapMap); JPT, Japanese HapMap; YRI, Yoruba HapMap. NA, Not applicable.

Predicted nonsynonymy frequencies

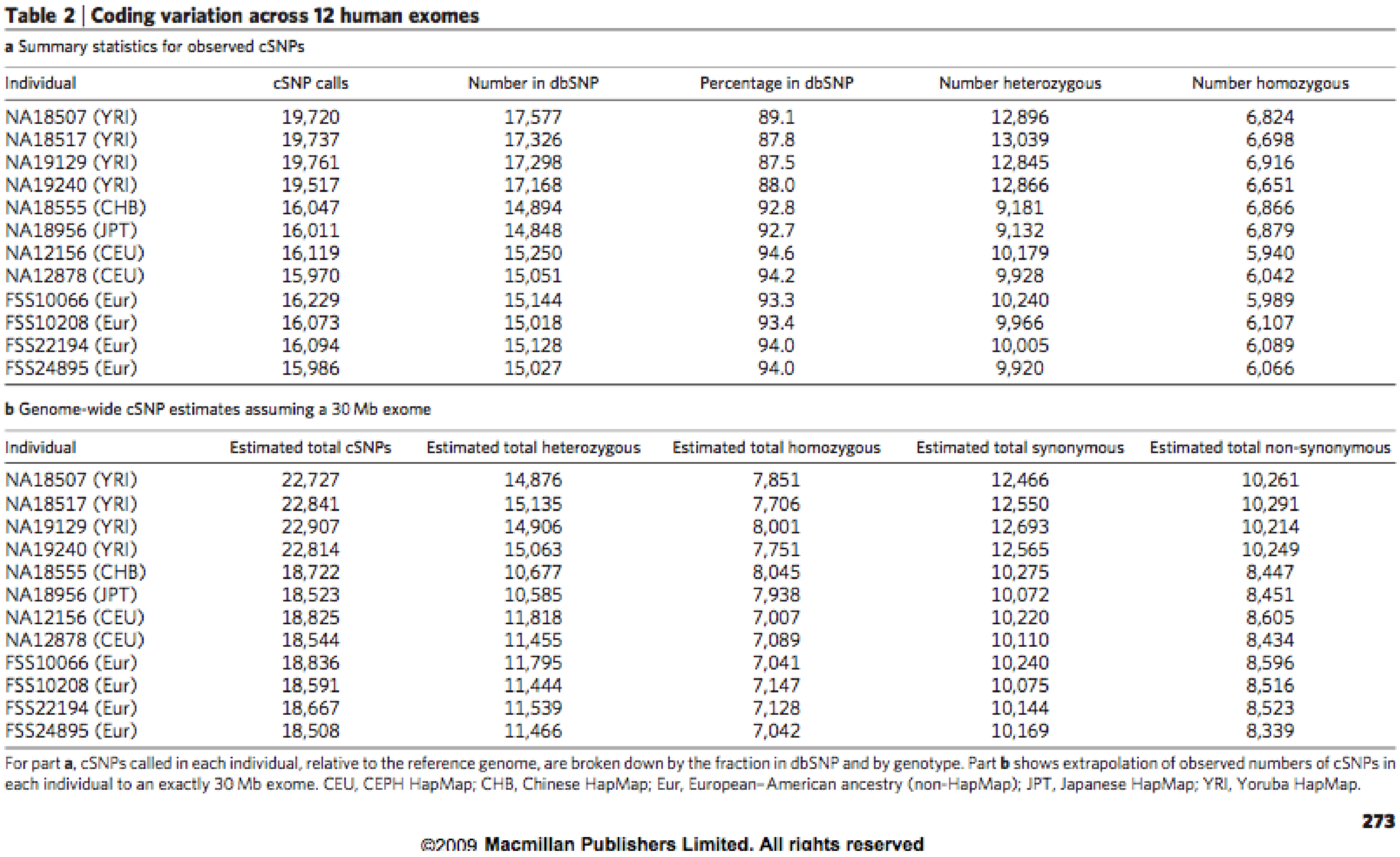


Table 2 | Coding variation across 12 human exomes

a Summary statistics for observed cSNPs

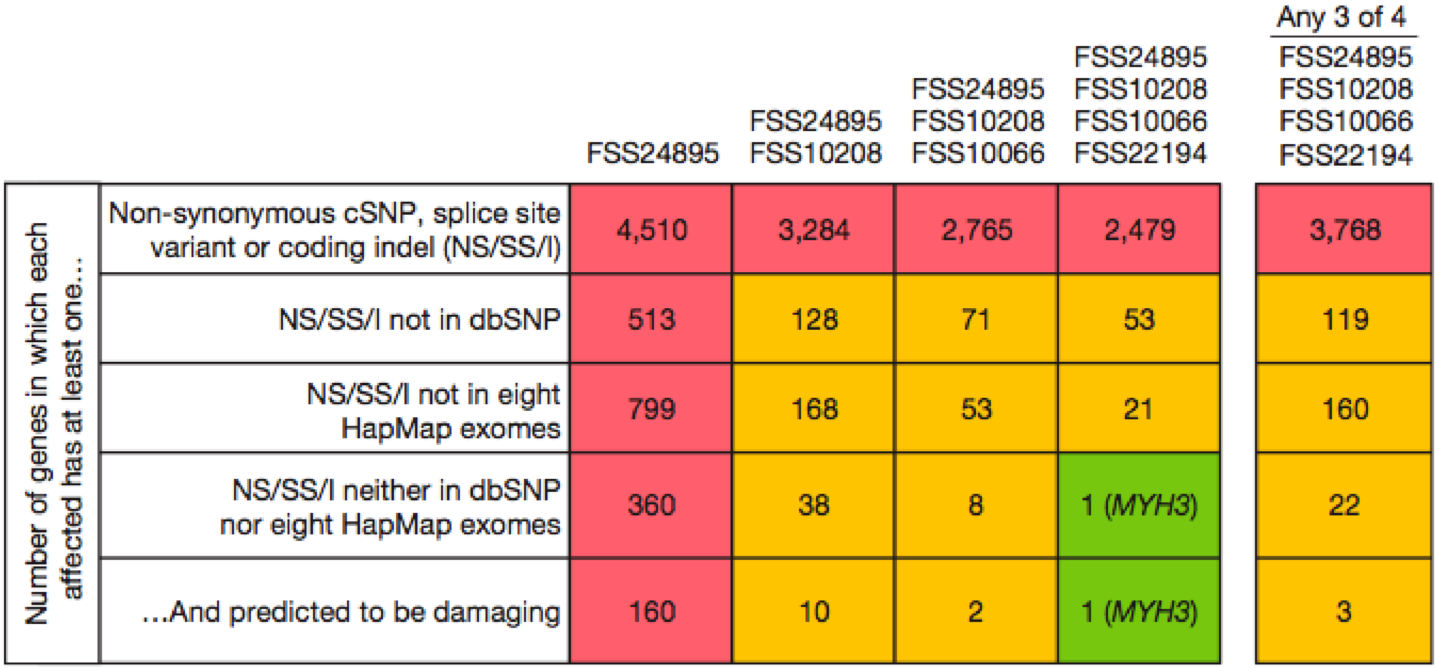
Individual	cSNP calls	Number in dbSNP	Percentage in dbSNP	Number heterozygous	Number homozygous
NA18507 (YRI)	19,720	17,577	89.1	12,896	6,824
NA18517 (YRI)	19,737	17,326	87.8	13,039	6,698
NA19129 (YRI)	19,761	17,298	87.5	12,845	6,916
NA19240 (YRI)	19,517	17,168	88.0	12,866	6,651
NA18555 (CHB)	16,047	14,894	92.8	9,181	6,866
NA18956 (JPT)	16,011	14,848	92.7	9,132	6,879
NA12156 (CEU)	16,119	15,250	94.6	10,179	5,940
NA12878 (CEU)	15,970	15,051	94.2	9,928	6,042
FSS10066 (Eur)	16,229	15,144	93.3	10,240	5,989
FSS10208 (Eur)	16,073	15,018	93.4	9,966	6,107
FSS22194 (Eur)	16,094	15,128	94.0	10,005	6,089
FSS24895 (Eur)	15,986	15,027	94.0	9,920	6,066

b Genome-wide cSNP estimates assuming a 30 Mb exome

Individual	Estimated total cSNPs	Estimated total heterozygous	Estimated total homozygous	Estimated total synonymous	Estimated total non-synonymous
NA18507 (YRI)	22,727	14,876	7,851	12,466	10,261
NA18517 (YRI)	22,841	15,135	7,706	12,550	10,291
NA19129 (YRI)	22,907	14,906	8,001	12,693	10,214
NA19240 (YRI)	22,814	15,063	7,751	12,565	10,249
NA18555 (CHB)	18,722	10,677	8,045	10,275	8,447
NA18956 (JPT)	18,523	10,585	7,938	10,072	8,451
NA12156 (CEU)	18,825	11,818	7,007	10,220	8,605
NA12878 (CEU)	18,544	11,455	7,089	10,110	8,434
FSS10066 (Eur)	18,836	11,795	7,041	10,240	8,596
FSS10208 (Eur)	18,591	11,444	7,147	10,075	8,516
FSS22194 (Eur)	18,667	11,539	7,128	10,144	8,523
FSS24895 (Eur)	18,508	11,466	7,042	10,169	8,339

For part **a**, cSNPs called in each individual, relative to the reference genome, are broken down by the fraction in dbSNP and by genotype. Part **b** shows extrapolation of observed numbers of cSNPs in each individual to an exactly 30 Mb exome. CEU, CEPH HapMap; CHB, Chinese HapMap; Eur, European-American ancestry (non-HapMap); JPT, Japanese HapMap; YRI, Yoruba HapMap.

273



		FSS24895	FSS24895 FSS10208	FSS24895 FSS10208 FSS10066	FSS24895 FSS10208 FSS10066 FSS22194	Any 3 of 4 FSS24895 FSS10208 FSS10066 FSS22194
Number of genes in which each affected has at least one...	Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,768
	NS/SS/I not in dbSNP	513	128	71	53	119
	NS/SS/I not in eight HapMap exomes	799	168	53	21	160
	NS/SS/I neither in dbSNP nor eight HapMap exomes	360	38	8	1 (*MYH3*)	22
	...And predicted to be damaging	160	10	2	1 (*MYH3*)	3

Figure 2 | Direct identification of the causal gene for a monogenic disorder by exome sequencing. Boxes list the number of genes with one or more non-synonymous cSNP, splice-site SNP, or coding indel (NS/SS/I) meeting specified filters. Columns show the effect of requiring that one or more NS/SS/I variants be observed in each of one to four affected individuals. Rows show the effect of excluding from consideration variants found in dbSNP, the eight HapMap exomes, or both. Column five models limited genetic heterogeneity or data incompleteness by relaxing criteria such that variants need only be observed in any three of four exomes for a gene to qualify.

Summary

- relatively intuitive filtering process leads directly to MYH3 as harboring more variants among FSS patients than controls
- Toydemir Nat Genet 2006

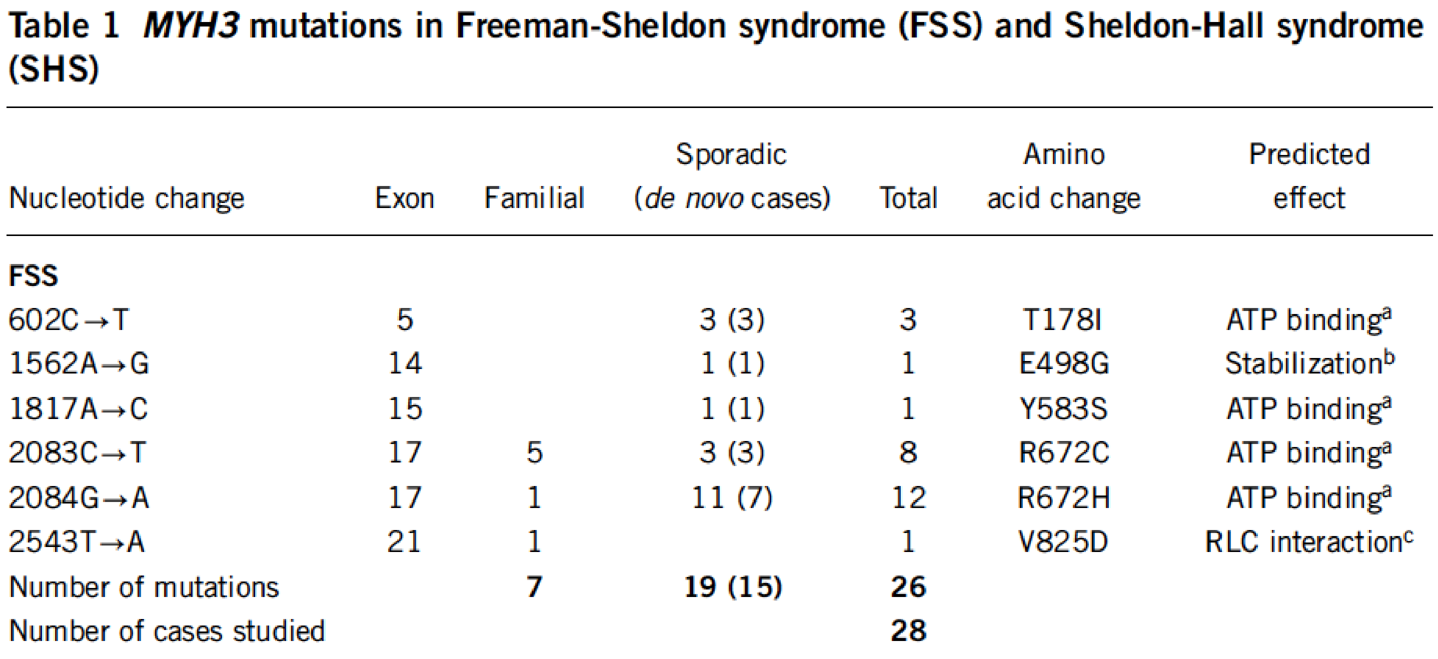
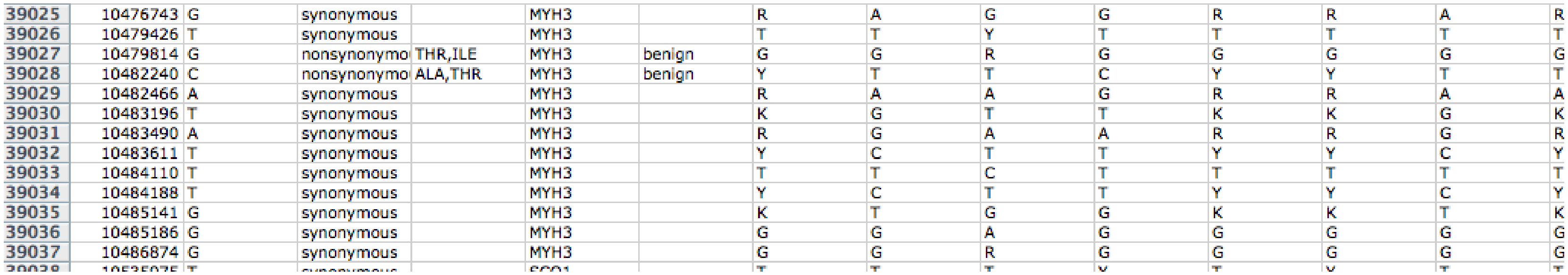


Table 1 ***MYH3*** **mutations in Freeman-Sheldon syndrome (FSS) and Sheldon-Hall syndrome (SHS)**

Nucleotide change	Exon	Familial	Sporadic (*de novo* cases)	Total	Amino acid change	Predicted effect
FSS						
602C→T	5		3 (3)	3	T178I	ATP binding[a]
1562A→G	14		1 (1)	1	E498G	Stabilization[b]
1817A→C	15		1 (1)	1	Y583S	ATP binding[a]
2083C→T	17	5	3 (3)	8	R672C	ATP binding[a]
2084G→A	17	1	11 (7)	12	R672H	ATP binding[a]
2543T→A	21	1		1	V825D	RLC interaction[c]
Number of mutations		**7**	**19 (15)**	**26**		
Number of cases studied				**28**		

Shendure distributes results on variants in MYH3 among controls

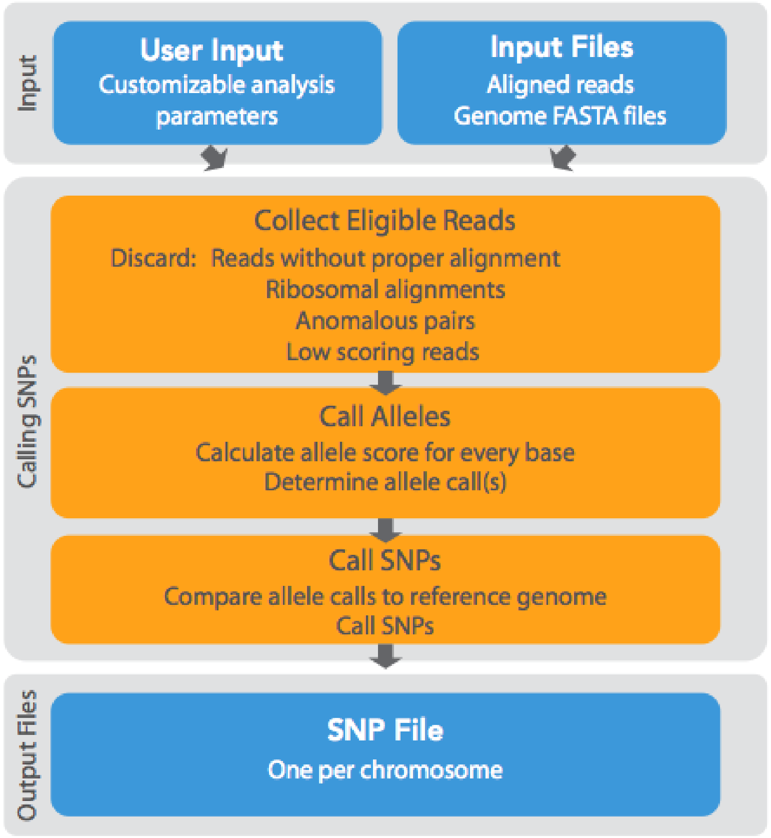


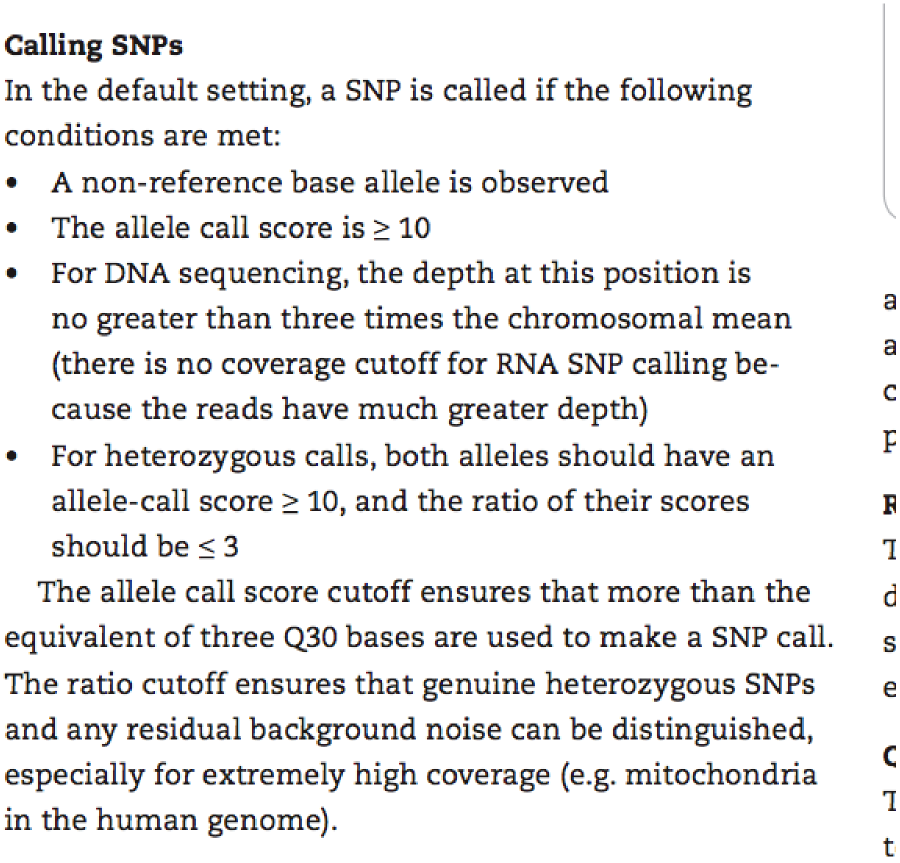
39025	10476743	G	synonymous		MYH3		R	A	G	G	R	R	A	R
39026	10479426	T	synonymous		MYH3		T	T	Y	T	T	T	T	T
39027	10479814	G	nonsynonymoı	THR,ILE	MYH3	benign	G	G	R	G	G	G	G	G
39028	10482240	C	nonsynonymoı	ALA,THR	MYH3	benign	Y	T	T	C	Y	Y	T	T
39029	10482466	A	synonymous		MYH3		R	A	A	G	R	R	A	A
39030	10483196	T	synonymous		MYH3		K	G	T	T	K	K	G	K
39031	10483490	A	synonymous		MYH3		R	G	A	A	R	R	G	R
39032	10483611	T	synonymous		MYH3		Y	C	T	T	Y	Y	C	Y
39033	10484110	T	synonymous		MYH3		T	T	C	T	T	T	T	T
39034	10484188	T	synonymous		MYH3		Y	C	T	T	Y	Y	C	Y
39035	10485141	G	synonymous		MYH3		K	T	G	G	K	K	T	K
39036	10485186	G	synonymous		MYH3		G	G	A	G	G	G	G	G
39037	10486874	G	synonymous		MYH3		G	G	R	G	G	G	G	G
39038	10535075	T	synonymous		SCO1		T	T	T	Y	T	Y	T	T

Exercise: Characterize variants in MYH3 (or another gene of your choice) in deeply sequenced HapMap individual NA19240. Pay attention to uncertainty of assertions of variant existence and type.

Technical considerations: Many approaches to calling variants

FIGURE 2: SNP CALLING WORKFLOW





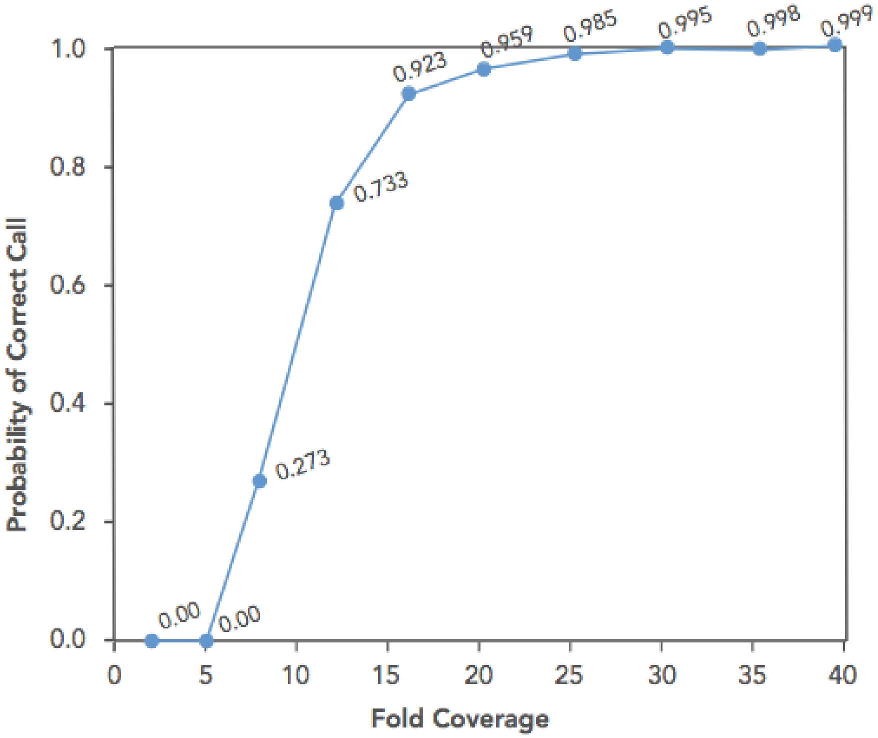
Calling SNPs

In the default setting, a SNP is called if the following conditions are met:

- A non-reference base allele is observed
- The allele call score is ≥ 10
- For DNA sequencing, the depth at this position is no greater than three times the chromosomal mean (there is no coverage cutoff for RNA SNP calling because the reads have much greater depth)
- For heterozygous calls, both alleles should have an allele-call score ≥ 10, and the ratio of their scores should be ≤ 3

The allele call score cutoff ensures that more than the equivalent of three Q30 bases are used to make a SNP call. The ratio cutoff ensures that genuine heterozygous SNPs and any residual background noise can be distinguished, especially for extremely high coverage (e.g. mitochondria in the human genome).

FIGURE 3: PROBABILITY OF CORRECT SNP CALL



Calculation of the probability of a correct SNP call at different coverage levels for a theoretical heterozygote position. The quality of the base calls was assumed at Q30.

SOAPsnp

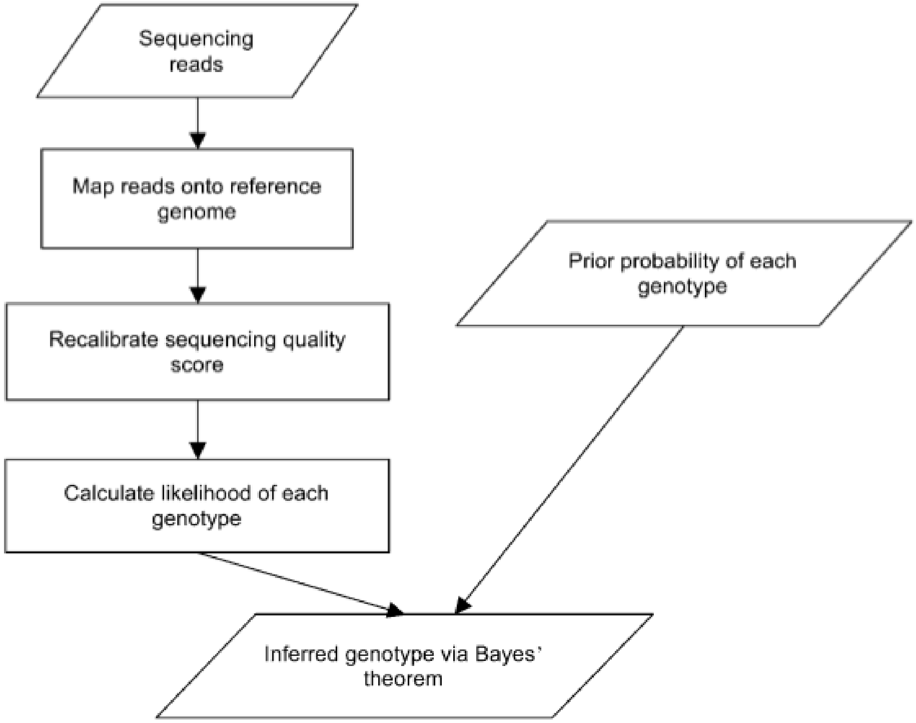

Sequencing reads
Map reads onto reference genome
Recalibrate sequencing quality score
Calculate likelihood of each genotype
Prior probability of each genotype
Inferred genotype via Bayes' theorem

Figure 1. Algorithmic overview of consensus calling for massively parallel resequencing. The program takes raw sequencing reads as input, maps them onto the reference genome, and calculates the likelihood of each possible genotype. It outputs the inferred genotype with highest posterior probability and its corresponding quality score.

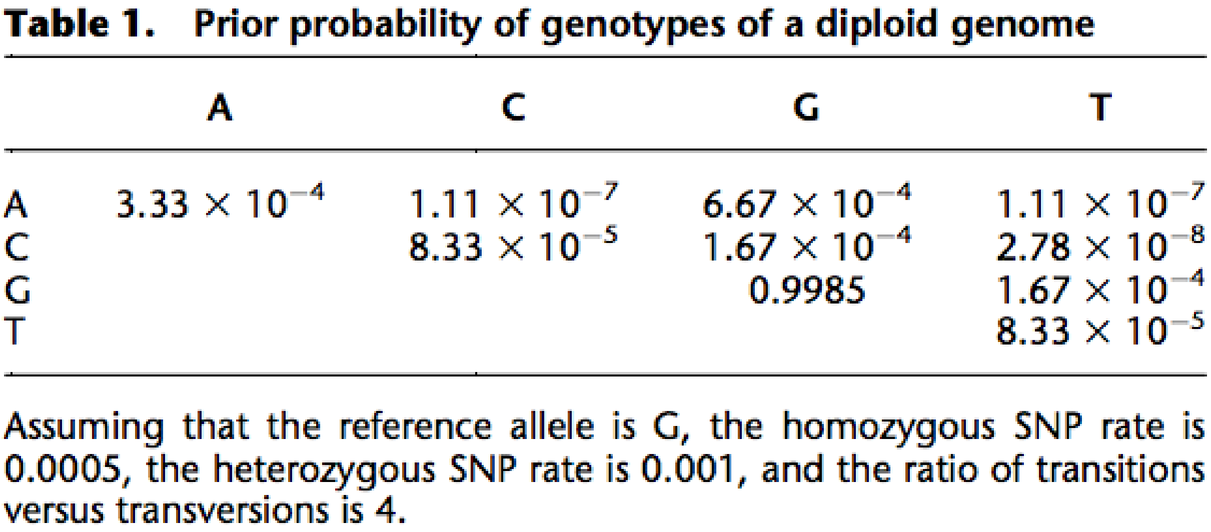


Table 1. Prior probability of genotypes of a diploid genome

	A	C	G	T
A	3.33×10^{-4}	1.11×10^{-7}	6.67×10^{-4}	1.11×10^{-7}
C		8.33×10^{-5}	1.67×10^{-4}	2.78×10^{-8}
G			0.9985	1.67×10^{-4}
T				8.33×10^{-5}

Assuming that the reference allele is G, the homozygous SNP rate is 0.0005, the heterozygous SNP rate is 0.001, and the ratio of transitions versus transversions is 4.

Exercise: Compare the results of your favorite NGS-based SNP caller with the Sanger-based 4mm hapmap phase 2 SNP genotypes for NA19240. Explain discrepancies.

Annotation and filtering resources

- Reference and individual genomic sequence:
  - consensus: BSgenome.Hsapiens.UCSC.hg18
  - SNP calls on 4mm HapMap phase II genotypes on 2 x 90 individuals: packages GGdata (CEU), hmyriB36 (YRI)
- Genomic features:
  - addresses of transcripts, exons: GenomicFeatures
  - addresses and assignments of dbSNP SNP: SNPlocs.Hsapiens.dbSNP.2008*
- Filtering:
  - Rsamtools for SAM/BAM formatted NGS data, interoperates with Bioc infrastructure very nicely
  - ShortRead+ for more general workflow components

Inference on rare variant existence and impact

- Inference and uncertainty: Not well-developed; ‘quality’ metrics are numerous
- Existence: mostly take for granted the information propagated by 1000 genomes, but we can compare with existing information on variants obtained via Sanger sequencing or SNP/CNV chips
- Impact: with R, plenty of tools ready to hand for matching, case-control testing and so on
- Representation problem: for SNP, currently focus on rare allele copy number, so a byte is more than enough, and some statistical procedures operating on raw bytes are available; indels and other complex variations need design considerations