

BEAGLE Imputation in SVS for Human and Animal SNP Data

0???1?1?01|1??1?0

1???|1?1?011??1?0

0???1?1?011??1?0

1???1?1?|011??1?0

January 11, 2017

Gabe Rudy
VP Product & Engineering



1 Overview Golden Helix

2 History of Genotype Imputation

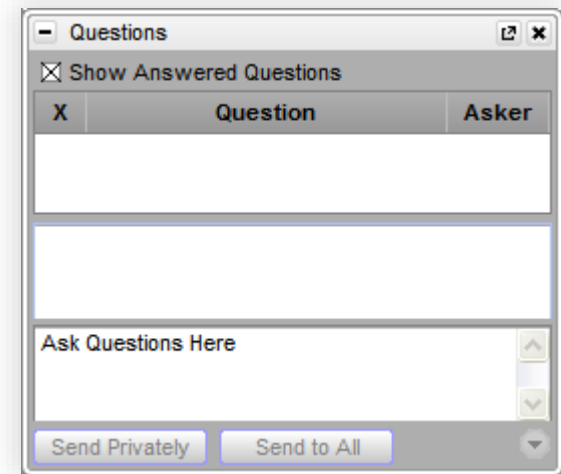
3 Why BEAGLE and Value of BEAGLE in SVS

4 Live Demo and Questions



Questions?

Use the Questions pane in your GoToWebinar window



Golden Helix – Who We Are



Golden Helix is a global bioinformatics company founded in 1998.



Filtering and Annotation
Clinical Reports
Pipeline: Run Workflows

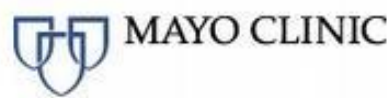


Variant Warehouse
Centralized Annotations
Hosted Reports
Sharing and Integration

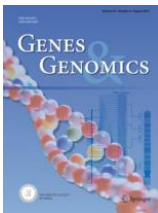
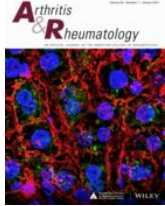
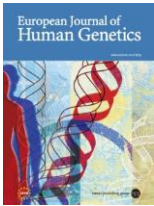
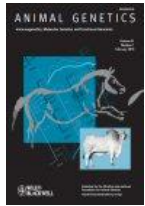


GWAS
Genomic Prediction
Large-N-Population Studies
RNA-Seq
CNV-Analysis

Over 300 customers globally



Cited in over 1000 peer-reviewed publications

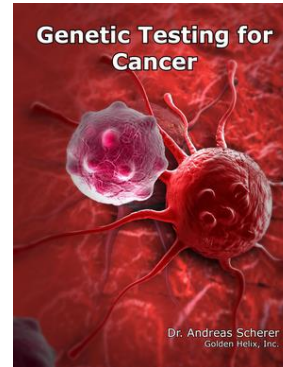


Golden Helix – Who We Are



When you choose a Golden Helix solution, you get more than just software

- REPUTATION
- TRUST
- EXPERIENCE



- INDUSTRY FOCUS
- THOUGHT LEADERSHIP
- COMMUNITY

- TRAINING
- SUPPORT
- RESPONSIVENESS

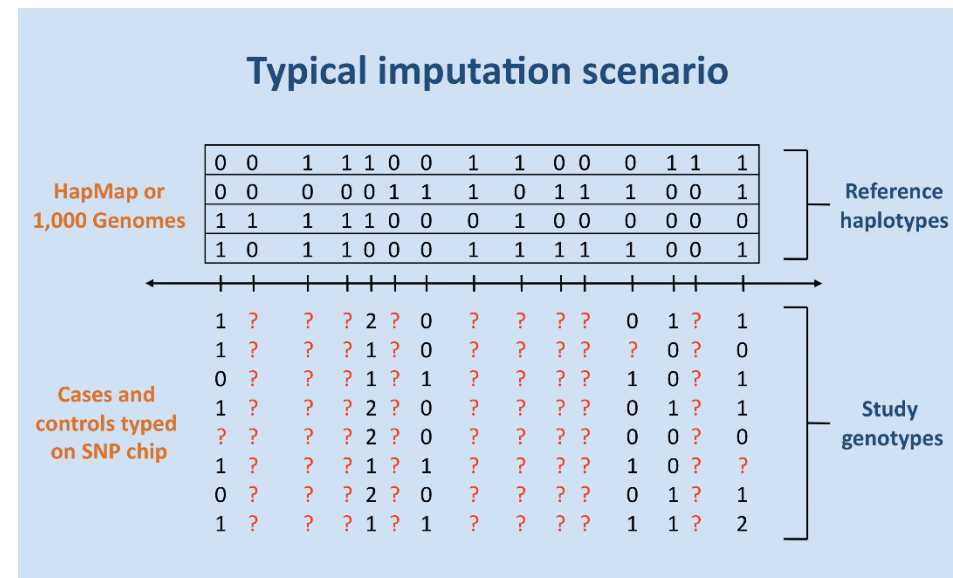


- TRANSPARENCY
- INNOVATION and SPEED
- CUSTOMIZATIONS

Why Imputation



- **Fill in Missing Genotypes**
 - Improve quality of GT calls
- **Harmonizing Arrays**
 - Facilitate meta-analyses that combine studies genotyped on different sets of variants
- **Increase Genotypes**
 - Increase the power and resolution of genetic association studies
 - Find candidate susceptibility variants to guide fine-mapping



Evolution of Imputation Methods



- **IMPUTE**

- June 2007

om/naturegenetics A new multipoint method for genome-wide association studies by imputation of genotypes

Jonathan Marchini^{1,2}, Bryan Howie^{1,2}, Simon Myers¹, Gil McVean¹ & Peter Donnelly¹

nature
genetics

- **BEAGLE v1**

- Nov 2007

Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering

Sharon R. Browning* and Brian L. Browning*

AJHG

- **BEAGLE v3**

- Feb 2009

A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals

Brian L. Browning^{1,*} and Sharon R. Browning¹

AJHG

- **IMPUTE2**

- June 2009

A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies

Bryan N. Howie^{1*}, Peter Donnelly^{1,2}, Jonathan Marchini^{1*}

PLoS GENETICS

- **IMPUTE2 (v2.3)**

- 2011 / 2013

- **BEAGLE v4**

- V4.0 - Dec 2015

- V 4.1 - Jul 2016

Genotype Imputation with Thousands of Genomes

Bryan Howie,^{*,1} Jonathan Marchini,^{*,1} and Matthew Stephens^{*,†}

^{*}Department of Human Genetics and [†]Department of Statistics, University of Chicago, Chicago, Illinois 60637, and [‡]Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom

Genotype Imputation with Millions of Reference Samples

Brian L. Browning^{1,2,*} and Sharon R. Browning²

AJHG



Value of Integrated BEAGLE

- **Integrated and Supported**
 - Don't need to run command line tools
 - Supported fully and integrates into the rest of your SVS analytics
- **Leverage SVS Data Management**
 - Import your SNP data from any source (PLINK, Illumina, Affy etc)
 - Also can import VCF
- **Handles NGS Variants as Well as SNPs**
 - BEAGLE 4.1 only reads VCFs with strict formatting requirements

Error while running BEAGLE for genotype imputation

Error while running **BEAGLE** for genotype imputation I am trying to run **BEAGLE** 4.1 for an imputation run. I have core exome chip **data** on variants of 20th chromosome in BED/BIM/FAM format, which I phased convert option in it. But, now when I try to run a **BEAGLE** imputation run by this: java -jar beagle.jar gt=test by hshabbeer.09

Phasing genotype per chromosomes in Beagle software

chromosomes in **Beagle** software I want to use **beagle** to phase my genotyped per chromosome, my **data** is in AB bgl missing=? out=Myrun , I am not getting phase **data** per chromosomes and I am also expecting the program by somakina

A: Background information and recommendations for phasing ASW whole genomes

You wouldn't need to re-phase the 1KG **data**, just tell **BEAGLE** to use it as a reference panel. by Zev.Kronenberg

Beagle 4.1 error : Possible data conversion issue

Beagle 4.1 error : Possible **data** conversion issue Hi, I have PLINK format **data** (PED/MAP) and I wanted to VCF so that I can input it in **BEAGLE** 4.1 to phase them, as **BEAGLE** only use VCF format. I wanted a trivial ran **beagle** (gt) on the input its giving me Java exceptions/errors. Its not a problem with **beagle** jar sample VCF format **data** downloaded from 1000Genomes. However, when I convert the **data** to VCF using PLINK PLINK and then use it as **BEAGLE** 4.1 input, then it doesn't like it. It'd be great if anyone can by aritra90

converting vcf to haploview with keep phasing

after haplotype phasing with **beagle** v4, now i have vcf (phased file **data**) file. but i don't know how phasing while converting. previous version of **beagle** software (v3.2) for haplotype inferring was good by goreishi

Phasing Data With Beagle

Phasing **Data** With **Beagle** Hey I need phased genotype **data** for another statistic I want to calculate and and I decided to phase my **data** with **BEAGLE**. Before even starting to phase I extracted with PLINK the can just phase this **data** set or hether youw ould recommend keeping the whole **data** set (all markers) but by Tim

BEAGLE 4.1 imputation

BEAGLE 4.1 imputation Hi all, I am new to the field of imputation. I am trying to use **BEAGLE** v4.1 to could be the target **data** for the imputation? 3. If I have a multiple target **data** sets, how can I do a a imputation of all of them with same reference **data** set at a time? 4. How to check the strand inconsistencies inconsistencies between reference and target **data**? If inconsistency occurs, how to make them consistent? Thank by cholingken

Strand Consistencies and SNPs vs Variants



- **Arrays Genotypes use Platform-Specific Allele Encoding**

- Illumina and Affymetrix defined their own “reference-independent” strand encodings

- **Sometimes Can Keep A/B Encoding**

- Different Arrays from Same Vendor

- **Sometimes Have Mapping to Human Reference**

- **Otherwise...**

Unsort		G 43467	G 43468	G 43469
Map	Sample	SNP_A-8427496	SNP_A-2179932	SNP_A-2207425
	Chromosome	1	1	1
	Position	164309029	164312048	164314302
	Cytoband	q24.1	q24.1	q24.1
	dbSNP RS ID	rs6672167	rs7524575	rs10800181
	Associated Gene	LOC284685	LOC284685	LOC284685
	Strand	-	+	-
	Strand Versus dbSNP	reverse	same	reverse
	Reference Alleles A/B	[A/T]	[A/G]	[C/T]
	Top Alleles	[T/A]	[A/G]	[G/A]
	Bottom Alleles	[A/T]	[T/C]	[C/T]
775	NA19722	A_A	A_B	A_B
776	NA19723	A_A	B_B	B_B
777	NA19724	A_A	A_B	A_B
778	NA19725	A_A	A_B	B_B
779	NA19726	A_A	B_B	B_B

Recode Genotype Column Data by Allele Name

Flip DNA strands for AGCT encoded genotypes

Transcode AB to AGCT encoding using mapping:

Marker map field in format 'A/B':

Reference Alleles A/B

Transcode using allele mapping:

Marker map field in format 'A:G B:T'

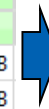
Chromosome



Recode SNPs to Variants

- **Select the RSID Map Field**
- **Select a dbSNP Annotation Track**
- **For each SNP we:**
 - Look up the SNP
 - Pull the allele frequency if present
 - Match your variants to the allele frequency, or the Major => Ref
 - Recode alleles (AB => AGCT)
- **Note this allows “lifting over” an older NCBI36 (hg18) snp data to GRCh37 (hg19)**
 - Also provides the Reference alleles required DNA-Seq analysis
 - Take advantage of new OMIM/CADD/OncoMD premium annotations

Unsort		G	12
Map	sub	SNP_A-4290489	
	Chromosome	22	
	Position	15268900	
	Cytoband	q11.1	
	dbSNP RS ID	rs5748616	
	Associated Gene	LOC100128190	
	Strand	-	
	Strand Versus dbSNP	reverse	
1	GSM233256_GSM233257	B_B	
2	GSM233258_GSM233259	B_B	
3	GSM233260_GSM233261	B_B	
4	GSM233262_GSM233263	B_B	
5	GSM233264_GSM233265	B_B	
6	GSM233266_GSM233267	A_B	
7	GSM233268_GSM233269	A_B	



Unsort		G	11
Map	Markers	SNP_A-4290489	
	Chromosome	22	
	Position	15268900	
	Reference	G	
	Alternates	C	
	Cytoband	q11.1	
	dbSNP RS ID	rs5748616	
	Associated Gene	LOC100128190	
1	GSM233256_GSM233257	C_C	
2	GSM233258_GSM233259	C_C	
3	GSM233260_GSM233261	C_C	
4	GSM233262_GSM233263	C_C	
5	GSM233264_GSM233265	C_C	
6	GSM233266_GSM233267	G_C	
7	GSM233268_GSM233269	G_C	

Creating a Reference Panel



- **Saves to local ImputeRefPanels folder**

- Saves as TSF, relocatable
- Uses current Project Genome

- **Allele Encoding**

- Recode to Reference/Alternate of reference sequence if possible
- If within same platform, alleles are matched alphabetically between reference and target samples

- **Imputed Data will use:**

- Column Headers
- Optional Map Fields

A screenshot of the 'Create Imputation Reference Panel' dialog box. The window title is 'Create Imputation Reference Panel' and it shows '565 samples and 131059 markers'. There are two tabs: 'Options' and 'Advanced'. The 'Advanced' tab is active. Under 'Reference Panel Output Options', there is a 'Folder' field set to 'ImputationRefPanels' with 'Reset' and 'Browse...' buttons. The 'Base Name' is '500K HapMap'. The 'Project Genome' is 'Homo sapiens (Human), GRCh37 hg19 (Feb 2009)'. The 'Allele Encoding' has two radio buttons: 'Alphabetically (A/B)' (selected) and 'Reference / Alternates'. The 'Included Map Fields' list includes: Cytoband, dbSNP RS ID, Associated Gene, Strand, Strand Versus dbSNP, Reference Alleles A/B, Top Alleles, Bottom Alleles, Flank, and Probe Count. At the bottom, there is a checkbox for 'Split Output by Chromosome' with a note 'Outputting by chromosome enables parrallele processing.' and a field for 'Number of concurrent runners (cores)'. The bottom buttons are 'Help', 'Restore Options', 'Save Options', 'Run', and 'Cancel'.

Running Imputation



- **Select from detected references**
- **Detects Allele Encoding**
- **Impute Regionally**
 - For targeted regions
- **Optionally output GT Probabilities**
 - Also drop low-prop GTs
- **Advanced**
 - BEAGLE Parameters
 - Trade off time vs accuracy

Genotype Imputation with BEAGLE

3500 samples and 388709 markers

Options **Advanced**

Reference Panel

Folder: [ImputationRefPanels](#)

Project Genome Filter: Homo sapiens (Human), GRCh37 g1k (Feb 2009)

	Name	# Samples	# Markers	Modified	romosom
1	chr22	2504	?		22
2	Small Panel	181	1355	2017-01-06	22

Only impute to ref markers within bp of target markers

Output

Base Name:

Spreadsheet as child of: Project Root Current Spreadsheet

Output Spreadsheet with per-Genotype Probabilities

Set genotype to missing if genotype probability is less than

Split Output by Chromosome

























Outputting by chromosome enables parrallele processing.

Number of concurrent runners (cores):

1000 Genomes Reference Panel



- **85 Million Variants in Phase3**
 - 2504 Samples
- **Extremely expensive to phase**
- **BEAGLE v4 Pre-Phased**
 - Per-Chr VCF files
 - Place in ImputeRefPanels folder
- **Use Regional Window!**
 - Use option

Name	Date modified	Type	Size
 chr1.1kg.phase3.v5a.vcf.gz	1/5/2017 2:03 PM	GZ File	737,074 KB
 chr1.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:00 PM	TBI File	205 KB
 chr2.1kg.phase3.v5a.vcf.gz	1/5/2017 2:04 PM	GZ File	788,570 KB
 chr2.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:00 PM	TBI File	217 KB
 chr3.1kg.phase3.v5a.vcf.gz	1/5/2017 2:04 PM	GZ File	671,966 KB
 chr3.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:01 PM	TBI File	179 KB
 chr4.1kg.phase3.v5a.vcf.gz	1/5/2017 2:04 PM	GZ File	690,066 KB
 chr4.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:01 PM	TBI File	173 KB
 chr5.1kg.phase3.v5a.vcf.gz	1/5/2017 2:47 PM	GZ File	599,745 KB
 chr5.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:44 PM	TBI File	162 KB
 chr6.1kg.phase3.v5a.vcf.gz	1/5/2017 2:49 PM	GZ File	624,556 KB
 chr6.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:44 PM	TBI File	154 KB
 chr7.1kg.phase3.v5a.vcf.gz	1/5/2017 2:49 PM	GZ File	557,489 KB
 chr7.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:45 PM	TBI File	143 KB
 chr8.1kg.phase3.v5a.vcf.gz	1/5/2017 2:49 PM	GZ File	520,870 KB
 chr8.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:45 PM	TBI File	131 KB
 chr9.1kg.phase3.v5a.vcf.gz	1/5/2017 2:49 PM	GZ File	408,758 KB
 chr9.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:45 PM	TBI File	109 KB
 chr10.1kg.phase3.v5a.vcf.gz	1/5/2017 2:49 PM	GZ File	475,940 KB
 chr10.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:45 PM	TBI File	121 KB
 chr11.1kg.phase3.v5a.vcf.gz	1/5/2017 3:03 PM	GZ File	466,314 KB
 chr11.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 3:00 PM	TBI File	121 KB
 chr12.1kg.phase3.v5a.vcf.gz	1/5/2017 3:03 PM	GZ File	453,060 KB
 chr12.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 3:00 PM	TBI File	120 KB



- **GWAS Follow Up**
- **Harmonize Cases and Controls**
- **Animal Genomics**





GOLDEN HELIX SNP & VARIATION SUITE

[Demonstration]

Upcoming in SVS 8.7



- Recode SNPs to Variants
- BEAGLE Genotype Imputation
- PhoRank Gene Ranking
 - Phenotypes are linked to genes through HPO and GO ontologies
- Various Polish Items

SVS

File Tools Download Resources Help

GOLDEN HELIX
SNP & VARIATION SUITE

Online Tutorials
Create New Project
Open Existing Project

- SVS_Bovine_SNPdata5
- PCA T1D and HapMap
- SNP_GWAS_Tutorial
- SNP_GWAS_Tutorial
- SNP_GWAS_Tutorial
- BEAGLE Test
- T1D SNP
- SVS8_MouseGP
- SVS_Illumina_BovineHD_SampleReadin
- PCA T1D and HapMap

Log Out Gabe Rudy

SVS 8.6.0 Release Notes

****DNA-Seq > Variant Classification**** has been replaced by the new **DNA-Seq > Annotate and Filter Variants** tool. (See [Annotate Variant Effect on Transcripts](#) for further details.)

New Features

- New **Secure Annotation** sources available:
 - CADD variant scores. (See [CADD](#) for more information)
 - OMIM Genes, Phenotypes and Variants. (See [OMIM](#) for more information)
 - MedGenomes's Oncology Mutation Database (OncoMD). (See [MedGenome OncoMD... Read more](#) »)

2016-10-18 16:39:37

PAG XXV – Heading to sunny San Diego!

This Saturday, Plant & Animal Genome (PAG) XXV will kick off in sunny San Diego! Here in Montana, we have had a brutally cold winter thus far. I, like many of you, am

support@goldenhelix.com, ph: +1.406.585.8137 or +1.888.589.4629

Version 8.6.0 Internal Win64 Released 2017-01-11

Active SNU with PBAT Analysis, Imputation, and Secure Annotations Expires 2018-01-01

© 2017 GoldenHelix



Questions or more info:

- Email info@goldenhelix.com
- Request an evaluation of the software at www.goldenhelix.com

