Q: Why haven't you responded to my e-mail?

A: We go out of our way to respond promptly to queries about **IMPUTE2**. If you wrote to us and haven't heard back yet, the most likely reason is that we are too busy to reply immediately. There are a few things that you can do to improve the chances of receiving a fast response:

- As you can see, we have put a lot of effort and accumulated wisdom into this website; please take a moment to see if your question is already answered in the documentation. We can tell when people have contacted us without reading the manual.
- Please be as specific as possible in your question (this will be easier if you've read the documentation). Sometimes open-ended questions are unavoidable, but these naturally take longer to answer.
- If the program is doing something you don't understand (e.g., crashing), it will be much easier for us to diagnose the problem if you send a copy of the screen output with your e-mail. Conveniently, **IMPUTE2** automatically writes an output log file, so you can just attach this file to your message.
- We are only human, and sometimes e-mails slip through the cracks, especially during busy times or holidays. If you haven't heard back from us after a week or so, feel free to e-mail again to check on the status of things -- we really do appreciate periodic reminders

Q. What is the strand issue in imputation and how can I deal with it in my study?

A. Genotypes produced by genotyping technologies and calling algorithms are “expressed” relative to either the + or - strand of the human genome reference. For example, a SNP which has alleles A and G relative to the + strand has alleles T and C relative to the - strand. Thus, it is crucial that each SNP has its genotypes expressed relative to the same strand in all the datasets used in imputation. All of the HapMap panels and 1000 Genomes panels have had their SNP alleles expressed relative to the + strand of the human genome reference sequence and so genotype data from study samples will also need alleles expressed relative to the + strand.

Possibly the easiest way to deal with strand is to re-orient your set of genotypes upfront, before phasing and imputation. Our program GTOOL has options that can re-orient a set of genotype data using a file that lists the strand (+ or -) of each SNP. Once this is done then you don't need to worry about it again.

Alternatively, if you have already pre-phased your data then you can deal with strand during imputation in IMPUTE2. We provide a number of options to allow you to deal with this issue

https://mathgen.stats.ox.ac.uk/impute/strand_alignment_options.html

IMPUTE2 will automatically align the strand between panels whenever it can do so

unambiguously; e.g., flipping A/C in Panel 2 to match G/T in the reference. The options on the webpage deal with variants where this is not possible i.e. A/T or G/C SNPs. If you have a file that lists the strand (+ or -) of each SNP the this file can be used to re-orient SNPs.

This is a very useful webpage that contains many of the strand files of commonly used chips.

http://www.well.ox.ac.uk/~wrayner/strand/

Q: What **`-Ne`** value should I use when there is more than one population in the reference panel?

A: The quick answer is ~20000. Modern imputation analyses typically involve reference panels with greater ancestral diversity, which can make it hard to determine the "ideal" **`-Ne`** value for a particular study. Fortunately, we have found that imputation accuracy is highly robust to different **`-Ne`** values; within each of several human populations, we have obtained nearly identical accuracy levels for values between **`10000`** and **`25000`**. We suggest setting **`-Ne`** to **`20000`** in the majority of modern imputation analyses (this will be the default value in future releases of **IMPUTE2**).

Q. What is the impact of including related samples in the study when carrying out phasing and imputation.

A. We think that including some related samples in the dataset can actually help with the phasing and imputation. In isolated populations methods have been proposed that actually try to uncover related samples and use them explicitly for phasing and imputation (Kong et. al. 2008 *Detection of sharing by descent, long-range phasing and haplotype imputation*. Nature Genetics http://www.nature.com/ng/journal/v40/n9/abs/ng.216.html).

Q. I am trying to use the windows version of IMPUTE. When i unzip the files and click on the .exe file a black small window appears and disappears in a few seconds, not giving enough time to read anything from it. This is followed by the generation of two files Test.impute2_summary and Test.impute2_warning.

A. IMPUTE2 is a command-line based program, so you need some kind of terminal window to enter the text commands that run the software. All mac and unix computers come with terminal programs, but you usually need to install one on Windows. I think most people do this via a program called 'cygwin': http://www.cygwin.com/

Once you have a functioning terminal window, you should be able to run the program using the kinds of example commands shown on our website.

Q. When imputing from the latest release of the 1000 Genomes Project haplotypes why does the progress of the IMPUTE2 run seem to slow down after 10 iterations?

A. The default behaviour of IMPUTE2 is to use 30 iterations. The first 10 iterations are burnin iterations where all IMPUTE2 is doing is phasing the study samples. After that IMPUTE2 starts to alternate between phasing study samples and imputing from the untyped SNPs from the reference panel. The latest release of the 1000 Genomes Project haplotypes have a VERY large number of SNPs and it takes a significant amount of time to impute them and causes a slow down in running time for iterations 11-30.

Q. After 3 iterations IMPUTE2 gives the message “RESETTING PARAMETERS FOR "EXTENDED FAMILY" MODELING”. Should i worry about this?

A. No, this is nothing to be worried about. At the beginning of the algorithm we run a few vanilla iterations to aid the mixing of the MCMC method. After those iterations we need to reset some internal parameters and this message is reporting that this is taking place.

Q. I downloaded the latest reference panel from http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated.html. However, I found that the genetic map files only have a total of 3,283,388 lines, while the .legend.gz files include a total of 40,504,230 SNPs (excluding chrX). So, are these genetic maps actually used, for a small proportion of SNPs?

A. Yes, the recombination maps are used. we use interpolation from the genetic to get the recombination rate between all consecutive pairs of SNPs.

Q. For imputation done in chunks of 5Mb, would it be OK for the last chunk to be bigger than 5Mb? Let’s say if the total imputation region is 5.5Mb, I guess the second chunk would be too small if I have to split it into two chunks. So, in my perl script, I cut all other chunks into 5Mb, except the last one with the 5Mb plus whatever is left. That could potentially make the last chunk bigger than 7Mb.

A. There is no hard restriction on using 5Mb. Our haplotype selection method works well when applied locally and we've tested it mostly with chunk sizes between 2Mb and 10Mb, although performance doesn't drop off fast at all. It is remarkably robust.

Q. Why is the total number of SNPs in the final .gen file slightly different from that in the reference panel .legend file. I think they should be exactly the same?

A. There might be SNPs in the input gen file that are not in the reference panel, but get included in the output, so the number of SNPs is not the same as the number in the reference panel.

Q. How do i encode missing genotypes in a gen file and how can IMPUTE2 be used to predict these genotypes?

A. Genotypes in a gen file are specified using 3 probabilities of each genotype. When IMPUTE2

reads a gen file it takes the genotype with the maximum probability IF it is above the calling threshold (defined using the option -call_thresh that has a default value of 0.9). All other genotypes are classified as missing. The default behaviour of IMPUTE2 is that it will not change the genotypes in a given dataset (i.e. in the file given using the -g option). There are two options that can modify this

-pgs - "Predict Genotyped SNPs": Tells the program to replace the input genotypes from the **-g** file with imputed genotypes in the **-o** file (applies to *Type 2* SNPs only). This option will replace ALL the genotypes with imputed genotypes, regardless of whether they are missing or not. This option can be useful if you want to check the overall quality of your dataset i.e. impute all chip SNPs using flanking information and see how this agrees with your original dataset.

-pgs_miss - Unlike **-pgs**, which replaces all input genotypes with imputed genotypes, this option tells the program to replace only the missing genotypes at typed SNPs. That is, any input genotype whose maximum probability exceeds the **-call_thresh** will simply be reprinted in the **-o** file, whereas input genotypes that fall below the calling threshold will be imputed in the output. In general, we think this option can work well if the levels of missing data are not very high, but it is possible that imputing a large amount of missing data at chip SNPs could cause subtle biased in downstream analysis.