**SUPPLEMENTARY MATERIAL 2**

**Guide to using the HPS-MPS pipeline**

**STEP 1: Setting up Docker**

This pipeline can be used on any platform once the docker container (virtual machine) has been installed in your system. For steps on how to install docker, go to the below link.

<https://www.docker.com/get-started>

or

<https://docs.docker.com/install/overview/>

You will see install options for Mac, Windows and Linux.

Mac – https://docs.docker.com/docker-for-mac/install/

Windows – https://docs.docker.com/docker-for-windows/install/

Linux (Ubuntu) - https://docs.docker.com/install/linux/docker-ce/ubuntu/#install-docker-ce

There are particular systems requirements for these installs. If you are having trouble installing the Windows software, please go to this link to install the docker toolbox (supports more Windows operating systems).

<https://docs.docker.com/toolbox/toolbox_install_windows/>

All links show you how to install and how to check you have installed correctly but doing a dummy run of

Docker run hello-world

Once you have docker installed and up and running. You will need to adjust the memory dedicated to the virtual machine so that the processes required to successfully run the hps-mps pipeline will be supported in the environment. This can be easily done using the user interface as seen below.

Please increase the memory to at least 16gb if possible. Increasing CPU’s will also shorten run time but that is up to the user.

If you have installed using the Windows toolbox option. You will need to run the following

**Change default vm settings**

If the default Virtual Box VM does not provide enough resources to give a good experience, we recommend you create a new VM with at least 2 CPUs and 16GB of memory.

* Double click the **Docker Quickstart icon from your desktop** and then run the following commands in that terminal.
* Remove the default vm

docker-machine rm default

* Re-create the default vm
	+ Choose the number of cpus with **--virtualbox-cpu-count**. For this example we'll use two.
	+ Choose the amount of RAM: **--virtualbox-memory**. This is also based on the host hardware. However, choose at least 16GB If you can.
	+ Choose the amount of disk space: **--virtualbox-disk-size**. It is recommended that this be at least 50GB since building generates a lot\* of output. In this example we'll choose 50GB.
	+ Create vm with new settings

docker-machine create -d virtualbox --virtualbox-cpu-count=2 --virtualbox-memory=16384 --virtualbox-disk-size=50000 default

* Restart docker

docker-machine stop exit

Then open a new Docker Quickstart Terminal.

**STEP 2: Setting up the hps-mps pipeline environment**

run the following command in your docker terminal

docker run -it -v ~/Desktop:/Desktop suswalsh/hpsmps

This will set up your environment and give it access to your desktop. Minimize this window unstill needed in step 4.

**STEP 3: Downloading folders needed for hps-mps pipeline**

1. Download the hps.zip file from here using your internet browser

<http://hirisplex.erasmusmc.nl/hps/hps.zip>

1. Unzip the file using whatever unzipping software you have on your computer.

3. **IT IS VERY IMPORTANT** to place the ‘hps’ folder on the **desktop** of your computer (make sure it is the folder itself and not the zipped folder. Its location should be ~/Desktop/hps/ and not ~/Desktop/hps/hps

**STEP 4: Preparing for your sequence files and running the pipeline**

1. Open the docker terminal window again, and run the following commands in this terminal window

chmod +x /Desktop/hps/scripts/\*.sh

/Desktop/hps/scripts/1-prepare.sh

If you get no errors at this stage, you have installed docker and environment correctly. If you have an error, please check to make sure the hps folder is on the desktop of your computer in the correct location (see above).

2. Minimize this terminal window and go back to your desktop hps folder

3. Now place your sample fastq files (any sequencer) into the hps/runfolder on your desktop, make sure they are **unzipped** (do not have .gz) before placing in hps/runfolder or immediately after placing in folder. **Do not proceed with the pipeline with .gz files.**

4. Edit the dirlist.txt file that is in that hps/runfolder (please use software that can save the file as a unix file – BBedit for Mac OSX or NotePad++ for Windows) to include your sample names (please use names up to the first \_ in the filename) as follows

sample1-of10-of-100\_S1\_L001\_R1\_001.fastq would be titled sample1-of10-of-100, there should be a sample name in each line as follows

sample1-of10-of-100

sample2-of10-of-100

etc.

Using your sample sheet from your MiSeq or Ion Torrent run is a good way to keep track of your sample names going into the pipeline. \*\*Currently the pipeline is set to running from 1-96 samples at once. If you include more than 96 samples, only the first 96 samples in the dirlist.csv file will run.

5. Go back to your terminal window that you left open and paste in the following commands one at a time (paste, and hit return, etc.)

/Desktop/hps/scripts/2-organisefiles.sh

/Desktop/hps/scripts/3-sampleinput.sh

/Desktop/hps/scripts/4-generatedata.sh

/Desktop/hps/scripts/5-makeonlinefile.sh

6. All results can be found in the hps folder - with the final result folder including the upload files for the prediction web tool. A description of every folder (within the hps folder) and what it contains after the pipeline has been run is described below.

 

POST –RUN OUTPUT

**Indelcheck** folder contains every samples check for the Indel variant of HPS (variant 1 in HP: rs796296176). There is only one indel in the HPS variants. Each sample is checked for the presence of this indel and a csv file is generated either if it is present or absent.

**Pipelinefiles** folder contain all the necessary files for using the pipeline. There are no results generated in this folder. Do not delete any files or adjust files in this folder unless you are experienced in doing so.

**Reference** folder contains the human reference files needed for alignment; in this case Hg19 is being used. Do not delete any files or adjust files in this folder unless you are experienced in doing so.

**Result** folder contains all the main result files of the run all summed up into single files. It includes the following files:



**The indelcheck file** is a complete list of all samples that had the variant 1 indel. IT IS VERY IMPORTANT TO CHECK THIS FILE as it influences the online and onlineupload file, if a sample shows the presence of the indel, then you must edit both the online and onlineupload file for that individual to show their correct genotype i.e.

If indel present in sample 123, go to sample 123 in the online.csv file and the onlineupload.csv file and change the genotype.

With no indel that variant would say CC (online file) or 0 (onlineupload file) for that sample, the user will have to edit this to say CA (if hetero) or AA (if homo insertion). Also for the onlineupload file which is used for prediction it should be changed from 0, to 1 or 2.

It is also best practice to edit the file found below in the TableFiles folder with this new genotype information (due to indelcheck). How to do this is described below.

**The online.csv** file is one file with every genotype result for all (up to 96 samples) that you ran.

**The onlineupload.csv** file is the online webtool file that you use to generate prediction probabilities for eye, hair and skin color using the website <https://hirisplex.erasmusmc.nl/>

This file needs no further editing (if the pipeline gave no errors and no indels were reported in the indelcheck file) and can simply be uploaded directly.



**runfolder** contains every file generated for each sample in each of their respective sample folder names.

I.e if you name a sample Sample123, there will be a folder called Sample123 in the runfolder. In here you will find:

Sample123.sam

Sample123.bam

Sample123sorted.bam

Sample123sorted.bam.bai

Sample123sorted.vcf

Sample123.recode.vcf

**Scripts** folder contains all the necessary scripts needed to run the pipeline. Do not delete any files or adjust files in this folder unless you are experienced in doing so.

**TableFiles** folder generates every samples .csv file which contains variant ref and alt alleles, genotype calls and read counts generated from the hps-mps pipeline using the programs as described in Figure 1 and the manuscript materials and methods. IF YOU HAVE AN INDEL IN A SAMPLE FILE, you must edit the samples genotype and read count for that variant 1 indel. See example on how below. Lets call the sample Sample123



This shows that this sample had an indel found at our variant 1 site. The most important part are the number of reads for the ref \* versus the +A at that site. Here it shows

COV = 1080

READS1 583: READS2 496.



The user **MUST** edit the Sample123.csv file found in the TableFiles folder as being Ref.forward of 583 and Alt.forward of 496 for that variant with ref being C and alt being +A

**\*\*\* THESE FILES ARE USED AS INPUT THAT MUST PASS THE THRESHOLD AND MIXTURE TOOL RULES FOR SAMPLE INTERPRETATION**

*\*\*\*In order to run the pipeline from scratch on new samples, it is recommended to re-download the hps folder and go from STEP 4 of this guide.*

TROUBLESHOOTING GUIDES:

If you experience any errors, some of the likely causes are listed below.

1. You did not enter the correct sample name into the dirlist file (or the file must be saved as a unix line breaks file)
2. Some of the fastq files contain no data (0 bytes) and hence that sample will not contain information in any of its generated files. If this happens you must delete the ‘empty’ .csv files in the **TableFiles** folder before running script 5 or it will give an error when making the final result folder files – *hence the online upload file will not be made.*