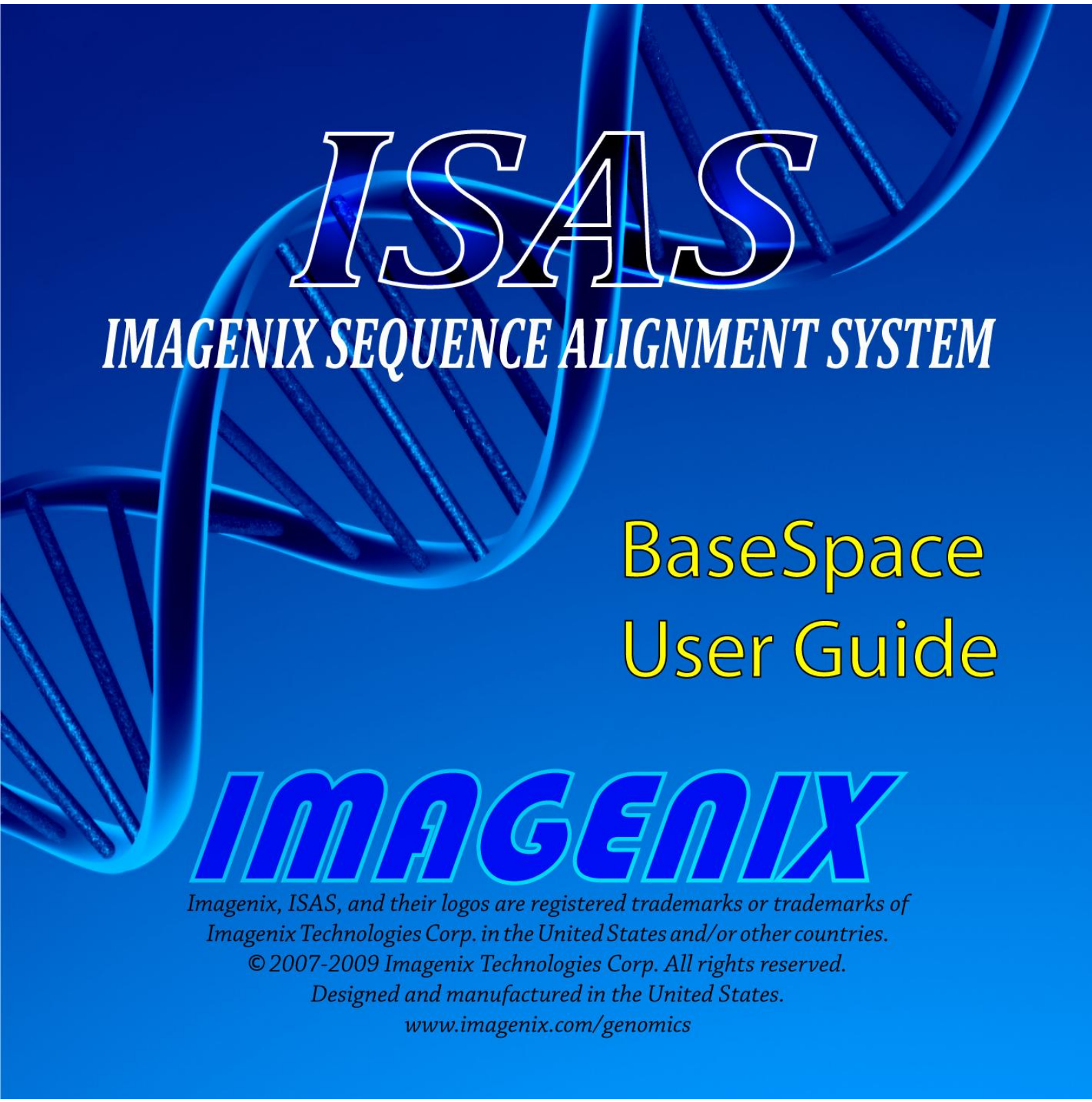

Table of Contents

I. Hardware Requirements

If you purchased ISAS as a software package to run on your own hardware, you should pay attention to this section. If you purchased a turn-key system (ISAS software pre-installed on an Imagenix computer) then you can skip this section.

1. CPU

ISAS is designed to run on x86_64 CPUs with, at a minimum, SSE3 capabilities. This includes Intel CPUs released since 2004 and AMD CPUs since 2005. You can confirm your processor's capabilities with your computer vendor. Imagenix can also supply you a simple program to interrogate your CPU for its capabilities. ISAS also reports the CPU capabilities it detects.

The installation DVD includes two executables. "ISASbasesNewCPU" will run on the latest generation CPUs (Intel Nehalem family or AMD Shanghai/Istanbul family). While "ISASbasesOldCPU" is for the older generations of CPUs. A computer with the older CPUs will refuse to run the "NewCPU" version, while the latest generation CPUs can run either versions of software, but will be usually be anywhere from 5 to 30 percent faster with the "NewCPU" version.

While ISAS will run on a system with merely a single CPU core, it is much more cost effective (compared with clusters of computers) to use a system with multiple cores. ISAS makes very efficient use of multiple CPUs and multiple cores. The constraint on the number of cores is the ability of the computer's memory controller and data bus to efficiently share the computer's RAM between all the CPU cores. At the time of writing of this guide, 8 cores (e.g. 2 CPUs with 4 cores each) are the most popular platform for ISAS (as well as most high performance applications), achieving almost 7 times the throughput of a single core system. Recently the cost and efficiency of hardware designs of 16 or more cores per computer have become much more attractive than one year ago, and we expect these more powerful systems to become more popular in the very near future. ISAS software is already designed to take advantage of this hardware architecture.

2. RAM

For small organisms, the memory requirement is usually trivial. But for larger reference genomes, we must pay attention to the memory requirements.

If your computer doesn't have enough RAM, and needs to use "virtual RAM" (also called "swapping", where hard disk is used to temporarily replace RAM) then your computer can come to a standstill. The minimum amount of RAM needed to run ISAS depends on the size of the reference genome. While a small organism like E.Coli with a 5 Million base (5MB) genome can easily be aligned on a laptop with only 2GB Ram, the human genome, with 3 billion bases (3GBase) requires a minimum of 16GB to run. The minimum requirement is 5bytes of RAM for each base of reference genome, and it is also a good idea to add roughly 1GB for overhead.

3. Hard Disk

ISAS requires approx. 16GB of disk space, plus the space for the input and output files. In the field of genomics/bioinformatics, is is not uncommon to encounter very large files.

Therefore, we recommend investing in fast storage systems. The input and output files can take up tens of GB of hard disk space per alignment run. A slow hard disk system, whether local or remote (NAS), can add 10 or even 30 more minutes to each alignment run. An off-the-shelf server ordinarily has a sustained read rate of 40MB/Sec to 100MB/Sec from its local hard disk. The Imagenix computer has a sustained disk read rate of 400MB/sec. We recommend you try to get the fastest storage rate possible when working with sequencing data files.

4. DVD Drive

ISAS software is usually supplied on a DVD-R disk, together with (as a convenience to our customers) the “hg19” public human genome reference. If your system has a DVD-R drive (which is common on modern computers) then you can easily load everything in a few minutes. Note: some computers have a DVD+R drive which cannot read DVD-R disks. Also, we found many Linux systems only allow “root” (superuser) to use the DVD drive. If your computer does not have a compatible DVD reader, but is connected through a network to another computer with a compatible DVD reader, you can read the DVD contents on the other computer in a few minutes, and copy them over your network in a few more minutes. In the case where you cannot read the DVD in any way, we can supply the software on a CD-R disk, or on a floppy disk, or on a USB flash drive, or by downloading from our web site. You can then download the reference genome over the internet (this can take an hour or more, depending on your connection bandwidth). Note that Linux has the tendency to assign “read only” attributes to files and directories which are copied from a DVD. Make sure to add write (“w”) privileges to the all the directories and files copied, and the execute (“x”) privilege to the executable files.

5. USB Port

In order to run, ISAS relies on a small hardware device, called a “dongle”, which needs to be plugged into the USB port on your computer. This is hardly a “hardware requirement” as virtually all computers made in the last ten years, from simple laptops to advanced servers, have a USB (Universal Serial Bus) port, and it is rare not to have many such ports on a modern computer.

II. Software Requirements

ISAS is available in Linux or Windows versions. In either case, a 64-bit operating system is required. The user must be able to use the USB port at a low level. Linux systems are usually configured in such a way that only “root” (system administrator) or “privileged” users (with system administrator powers) can access the USB port at a low level. Since ISAS uses a “dongle”, a device connected to the USB port, the Linux system should either be set up to allow every user to use the USB port, or the user should have “root” privileges. In addition, many Linux installations were done with “swappiness=60” which is possibly appropriate for desktops, but deleterious for server performance. The value of swapinnes should be set to zero. For both of these Linux configuration problems, see the “Linux Settings” section of “Installing ISAS” below.

III. Installing ISAS

If you purchased a turn-key system (ISAS software pre-installed on an Imagenix computer) then you can skip the “New Installation” section. You are already able to run alignments on the human reference genome. To add or switch to different species, read the section that follows it: “Adding or Changing Reference Genomes“.

1. New Installation

A. Linux Setting - Swappiness

Linux has a parameter called “swappiness”, which needs to be set to zero, otherwise the system will swap memory needlessly, slowing down ISAS. You can set swappiness to zero each time you run ISAS by typing (you need to be “root” or “superuser” to do this):

echo 0 > /proc/sys/vm/swappiness

You can always check the current swappiness value by typing (no need to be “root”)

cat /proc/sys/vm/swappiness

Or, a better alternative is to permanently set it to zero. As “root”, or “superuser”, edit the file

/etc/sysctl.conf

In the above file, we need to add the following line:

vm.swappiness=0

If, for example, there is already a line such as

vm.swappiness=60

then instead of adding a new line, just change the “60” (in the example above) to “0”. A re-boot is required to load the new swappiness value from this file, after which it is still prudent to check (“cat /proc/sys/vm/swappiness”) that the value is truly “0”.

B. Linux Setting - Regular Users Accessing USB Port

When installing a new ISAS software on your own computer (as opposed to running a turn-key system preinstalled ISAS software on an Imagenix computer), we recommend that you first verify that you can successfully run as root, and only afterwards, if required, add the capability for regular (non root) users to access the USB port, enabling them to run ISAS. When you reach that stage, return to this section. Some administrators find it easier to use the “sudo” command to allow non-root users to use ISAS than to add privileges for non-root users to access the USB ports.

While different Linux distributions can have different mechanisms for this, below we describe the simple way to grant non-root users permission to access the USB device, which has been tested on Redhat Enterprise 4+, CentOS 4+, Fedora 5+, and SUSE 10+.

As root (or super user), go to the directory

/etc/udev/rules.d

There can be several files there. These define the rules (e.g. permissions) for using devices plugged in to the USB ports. The rules from the different files are applied in alphabetical order of the file names. Create the file

ISAS.rules

With the following content

SYSFS{idVendor}=="07f2", SYSFS{idProduct}=="0001", MODE="0666"

A re-boot is required to take effect, and then non-root users can access the USB device.

C. Begin Installation

ISAS came with a DVD (disk) and a dongle (small USB device). Insert the disc into your CD/DVD reader, and insert the dongle into one of your USB ports. Make sure the dongle is fully inserted, and the connection is stable.

Create a directory on your hard disk for ISAS to reside in, for example "/home/ISAS-bases". We will use this example in our description, but you can use any other valid directory name. Copy the contents of the ISAS DVD to "/home/ISAS-bases", preserving the sub-directory file structure. The directory "/home/ISAS-bases/data" will be used to build the reference database, and its subdirectory "/home/ISAS-bases/data/reference" will hold the reference genome.

Important Note: After copying from the DVD to your hard disk, some Linux systems may arbitrarily set the permissions of files and directories (e.g. permission "r" as opposed to "rw", or maybe only useable by root). This may prevent ISAS from functioning properly. Make sure to enable read and write privileges ("rw") to the ISAS directory and its two subdirectories ("data" and "data/reference") as well as the files they contain - especially the file "settings-bases.txt". Also set "execute" privilege ("rx") to the executable file(s) ISAS*. These privileges should be set for whatever category of users will be using ISAS (e.g. root only, or all users).

For your convenience, the ISAS DVD already includes a compressed version of the public "HG19" human genome reference; therefore, if copied correctly, the directory "/home/ISAS-bases/data/reference" should now contain 50 files, 2 for each chromosome (1 through 24) plus 2 for mitochondrial DNA. The executable program file will be called "ISASbasesXX" where XX denotes the version, for example "ISASbasesNewCPU" for latest generation computers, or "ISASbasesOldCPU" for computers based on previous generation CPUs. You can always try to run ISASbasesNewCPU and it will tell you if your computer is not compatible, and then you know that you should be using the other version. Depending on your DVD read rate and hard disk write rate, copying should take a few minutes. We suggest first completing the installation and testing for the human reference genome, before changing or adding different species/reference.

Change directory to the ISAS directory (type "cd /home/ISAS-bases") and run ISAS (type "./ISASbasesNewCPU" or whatever the name of your executable is). Because this is the first time, ISAS will emit an error statement, complaining about a missing "bin" file, which is normal. However, if an error statement is "not installed correctly", then the USB dongle is not present, or your computer is unable to access the USB port (see section II above). If the error message is "reference not loaded properly", then you haven't copied all the files from the DVD into the "/home/ISAS-bases/data/reference" directory.

For each reference, a bin file is required. We will now create a bin file for the hg19 reference which is in the “data” directory. Type, “MAKEBIN”, and then press the ENTER key. ISAS will ask “Are you sure ?”, at which point you should type “YES” and press the ENTER key once more. ISAS will now prepare the optimized database for the current reference. This process takes between 5 to 15 minutes, depending on your computer, and consumes about 13GB of disk space. When finished, exit ISAS by typing “EXIT” or “QUIT” (followed by the ENTER key). You’re ready to go to section IV. If at some later time you want to change the reference genome, you will need to run MAKEBIN again. If you forgot to give all the directories write (“”rw”) permission, the bin file will not be able to be saved without error.

2. Adding or Changing Reference Genomes

When ISAS is run, it expects the reference and its database to be in a specially structured database subdirectory. It will also create a settings file in the current database subdirectory, which stores your desired settings for using this reference. You can have many different reference genomes to work with, as long as each reference resides in its own directory, and is created and structured correctly. For example, to add “hg18” reference genome (in addition to the “hg19” reference):

a. Create a directory in the location where the ISAS executable resides called “hg18”. Create a subdirectory under “hg18” which is called “reference”. If we use the example, where ISAS was installed in the directory “/home/ISAS-bases”, then the new database directory will be “/home/ISAS- bases /hg18” and its subdirectory will be “/home/ISAS- bases /hg18/reference”.

b. Put the “fa” files of all the chromosomes in the “/home/ISAS-bases /hg18/reference” directory. The names should be “chr1.fa”, “chr2.fa”, etc. Only numbers are allowed, so, in our example, the files “chrX.fa”, “chrY.fa”, and “ChrM.fa” (if the mitochondrial DNA is also desired) should be renamed to “chr23.fa”, “chr24.fa”, and “chr25.fa” respectively. For E.Coli, for example, the entire genome should be named “chr1.fa”.

c. Run ISAS (if not currently running). Type:
DATABASE=hg18

d. This sets the current reference database to hg18. Send the command by ending with the ENTER key.

e. Set the number of chromosomes by using the “CHR” command. For example, for human with mitochondrial DNA, type “CHR=1,25”, without mitochondrial DNA type “CHR=1,24”. For E.Coli type “CHR=1,1”. The command is sent when you type the ENTER key at the end.

f. Type “FASTA” (followed by ENTER). ISAS will process the “fa” files. This should take a few minutes.

g. Use the MAKEBIN command for either standard, or valid adjacent, or both, as described in the above section III 1. This can take 1 to 10 minutes, depending on your computer.

Note: to switch back to hg19, you can type
DATABASE=hg19
To return to hg18 use the command
DATABASE=hg18

IV. Running ISAS

We recommend that the first time you run ISAS on *your own computer* (i.e. not a turn-key Imagenix system), you do so as "root". This is to avoid a common problem with Linux systems, where normal users are blocked from using the USB port. Only after you are satisfied that the installation is successful, and can run ISAS as root with no difficulties, we suggest that you try to make it useable by non-root users (if required). This is described in section III 1 b on page 5.

ISAS will make use of all of the RAM on your system. It is important not to run any other programs, and to disable unnecessary memory swapping. If you didn't permanently set the "swappiness" value to zero on your system (this is highly recommended), then you will need to remember to set it to zero just before running ISAS every time (see section III 1 on page 5 for instructions). Many Linux installations have a default value of 60 for swappiness, which causes the system to try to "predict" what programs need RAM. This can cause ISAS to slow down dramatically, making it almost useless in some cases. This is all preventable by setting swappiness to zero.

ISAS can be run in interactive mode, or in production mode. When ISAS is run for the first time with no input arguments, it enters interactive more. It will load the reference into memory, and await further commands. You can interactively issue different commands (one by one) to get information about your hardware, reference, and settings, or to change settings, or to perform alignment on files. When you quit your interactive session, the settings are saved, to be reloaded the next time ISAS is run (either in interactive or production modes).

In production mode, ISAS is run (from the operating system command line, or from your calling script) with a single argument which represents a command. ISAS will start, load the settings last saved, execute the single command from the command line, then save the settings (your command could have been to change some setting), and exit.

We suggest that you use interactive mode to get familiar with the different capabilities and characteristics of the system. At a later stage, you can use production mode in batch files or under control of an automated program.

Once in an interactive ISAS session, a command is issued to ISAS by typing one of the supported commands, often followed by "=" and then the list of parameters (if any), separated by commas. The ENTER key is pushed to begin processing the command. ISAS Commands are all case insensitive (except file names, which in Linux are case sensitive).

You can easily get a list of the known commands by typing "?", followed by the ENTER key. From this point on, we won't mention "followed by the ENTER key" anymore – it should be implicitly understood that a command must be terminated by ENTER.

This is a transcript of the above session:

```
Enter next command, or type "?" (and ENTER) for list of commands.
?
Enter one of the following:
```

```
1.  "STATUS" Display current system mode/parameters.
2.  "SEQUENCE=x" where x is ReadLength (currently 71) bases (each: A,C,T, or G).
3.  "FILE=x" Process fastq sequence input file x.
4.  "FILES=x,y,min,max" Process two fastq paired sequence input files x and y. min
    and max are the minimum and maximum base pair distances allowed between pairs.
5.  "MODE=x,y" Set ReadLength to x and allowed mismatches to y.
6.  "INDEL=x" x=0 for ungapped, x=1 for gapped, or x=2 for hybrid alignment.
7.  "VERBOSE=x" x="0" for Non-Verbose output, "1" for Regular, "2" for
    Unique-Only, or "3" for SAM output file format.
8.  "RANDOM=x" Create a file of x random 71mers from current genome.
9.  "RANDOMS=x" Create a pair of files of x random 71mers from current genome.
10. "LIMIT=x" For each sequence, stop searching after x matches are found.
11. "REF=x,y" View reference chromosome x position y.
12. "CHR=x,y" Set first chromosome to x and last chromosome to y. If no chromosomes use
    "CHR=1,1". For human, use "CHR=1,25" to include mitochondrial DNA, or "CHR=1,24".
13. "FASTA" Convert fasta files to reference files. Needs to be done once.
14. "MAKEBIN" Create database files from reference files. Needs to be done once.
15. "HARDWARE" Display information about system hardware.
16. "FILTER=x" x=0 for No Filter (100% sensitivity). 1 (min) through 10 (max) to
    filter sequences.
17. "DATABASE=x" Change reference database. Load it from the directory "x".
16. "QUIT" or "EXIT" to exit.

Enter next command, or type "?" (and ENTER) for list of commands.
```

1. STATUS

Use the STATUS command to find out the current settings. First, ISAS will report the first and last chromosomes (both 1 if no chromosomes), and total number of bases in reference. Next, ISAS will report the current Mode which includes the ReadLength, max. no. of mismatches allowed (see 5 below), and whether mismatches include indels (“gapped”) or not (“ungapped”) as described in 6 below. Next, Verbose or Regular output (see 7 below), and the value of LIMIT (see 10 below). Finally, the buffer size, which is the number of input sequences aligned at a time, will be reported (see section V 3 on page 17).

2. SEQUENCE=x

You can use ISAS to perform alignment on a single sequence, as opposed to a file with millions of sequences. With this command, ISAS will display a “mismatch analysis” between the input sequence and the reference, for each match found (up to LIMIT, which is the maximum number of matches searched for). An “x” will be shown for each mismatch, with spaces (gaps) filling the reference in case of insertions, or filling the sequence in case of deletions. This command is useful for understanding why a particular sequence is considered a match in a particular mode. The “x” in the “Sequence=x” syntax represents the sequence. The sequence should be composed of bases, where each base is either “A”, “C”, “T”, or “G” (case insensitive). If the length of the entered sequence x is less than ReadLength, as determined by the current settings, ISAS will reply with an error message indicating this problem. If x is too long, it will be truncated to the proper length, and a reply will contain a warning as to the truncation. Below is an example session transcript, the

mode was “MODE=50,2” which means that the Read Length was 50, and the total allowed mismatches was 2 (See section 5 below).

```
Enter next command, or type "?" (and ENTER) for list of commands.
sequence=ATGGAAAGCATCTCCGACTGTGGCAAATGTTCATAACCTCCTCCCTCCCC
One match found. 7.0 micro seconds.
Match no. 1: Forward Chr. 6 Positions 1000000..1000049, 2 Mismatches
 ATGGAAAGCATCTCCGACTGTGGCAAATGTTCATAACCTCCTCCCTCCCC
 ATGGAAAGCATCTCCGACTTTGGCAAATGTTCATCACCTCCTCCCTCCCC
                     X              X                 2 substitutions
Enter next command, or type "?" (and ENTER) for list of commands.
```

The sequence was found, exactly in one location in the human genome: chromosome 6, genomic position 1,000,000, forward strand, and with a total of 2 mismatches. The input sequence is shown directly above the reference at the matched location, and the two mismatches are indicated by “X” below them. If we were to run the same command in Mode=50,1 (Read Length=50 but Allowed Mismatches=1) the sequence would not have been found.

3. FILE=x

Use this command to perform alignment on a file of sequences. “x” denotes the file name, including the full path. The file should be in “fastq” format. In this format, each sequence is represented by four lines. The first line of each quadruple is the tag line. Usually this is the ID tag that the sequencer assigned to this sequence data. This line will be echoed in the output file, with the first character (usually a ‘@’) skipped. The second line will contain the sequence, while the third and fourth lines will be ignored.

The mapping results will show the original sequence followed by its tag on the first line, followed by one line for each match found - if any. The line will start with the reference bases in the match position, with the mismatched bases converted to lower case, followed by the position description. Forward strand matches will be denoted by a “+”, while “-“ denotes a reverse strand match.

If the length of the sequence is too long, as determined by the current Read Length, ISAS will ignore the extra bases. If it is too short it cannot be aligned, so you should take care not to try to align a file of 25mers as if they were 35mers. The following is an example of an input file.

```
@SEQUENCER102_25:4:1:1:329/1
TAGACAAACAAAAAACCTTCCTCTCAATCCTATCACTTGTAATTTAAGAA
+
"IIIIIIIIIIII31IIGBA@IDII?020;4IDE74>37,2/-00599.56
@ SEQUENCER 102_25:5:1:2:239/1
GACATTGGGAGGCTGAGCCAGGAAGATGGCTTGAGCCCCGGAGGTCAAGA
+
```

"IIIIIIII/III8I5I@'I2?=2-),/::-;+2.0+0#,4%-".+*),&+
@ SEQUENCER 102_25:5:1:2:2047/1
CAATAAATTAAACCTTATATCTTGCCAGGTGCAGTGGCTCATGCCTGTAA
+
"IIIGAII@G<1?DII8II7IIF0(:00%&/4.?&6+1/7+'0'1%,($#,
@ SEQUENCER 102_25:5:1:2:822/1
ACTTGGGCTGCTGTATGCGAAACTGGTGGCCCCTGAGGCCACGAAGCTAC
+
"2I@87II,IIIB.8DHD?4II4;42&/%204+2-,(&209#/*&0"0+2&

ISAS will create a name for the output file that will capture the important aspects of the settings used to perform the alignment. It will start with the name of the input file, and then concatenate the current read length and allowed mismatches. “Gapped” or “Ungapped” will indicate the current INDEL setting. Similarly the LIMIT and VERBOSE level will also be appended to the file name. Finally, “.txt” will be added at the end.

The location of the output file will be the same as that of the input file. Make sure you have enough disk space when running this command. Output files are usually much larger than their corresponding input files, as they contain the sequence and tag from the input file, plus the alignment results. For high LIMIT (maximum no. of matches) values, the output file size can become very much larger than the input file size, as potentially many non-unique match positions can be listed for sequences in repeat regions of the genome. Below is an example session transcript, the mode was 50,2 (see section 5 below), i.e. the Read Length was 50 and the Allowed Mismatches was 2. LIMIT (see section 9 below) was 5 and INDEL was disabled (Ungapped).

```
Enter next command, or type "?" (and ENTER) for list of commands.
file=/home/Data/file1.fastq

Created output file /home/Data/file1.fastq.ReadLength50-Mismatches2-Limit5-Ungapped-
Normal.txt (0.0 sec.)
Buffer cleared (0.3 sec.)
Aligning...

Aligned 53325569 sequences (659.1 sec.)
Wrote   53325569 sequences (231.1 sec.)

Total of 53325569 sequences done in a total of 14.8 Minutes

Results summary written to file: /home/Data/file1.fastq.ReadLength50-Mismatches2-Limit5-
Ungapped-Normal.Stats.txt

  Hits     Histogram     Percent
 ======  ============  =========
     0       3590892         6.7
     1      39572266        74.2
     2       1625012         3.0
     3        735439         1.4
     4        503033         0.9
     5+      7298927        13.7

Enter next command, or type "?" (and ENTER) for list of commands.
```

As an added convenience, a histogram of the mapping results is displayed at the end. It shows how many times sequences were found, from zero (sequence not found in reference) to LIMIT+ (sequence was found LIMIT or more times). This is very useful for immediately identifying a problem, such as a user error or a bad experiment. For example, if usually we expect about 70% of sequences to be uniquely found, then seeing almost all the sequences as not found immediately raises our suspicion, and we can look around quickly to find that we typed in the wrong Read Length, or maybe the wrong species. Without the histogram, we would have to analyze Gigabytes worth of data to detect these simple human errors. Another example, where the histogram shows far too many sequences as having large number of repeated matches, draws our attention and we find that the sample was not prepared properly for sequencing, resulting in data that is mostly A's.

This command will be used to illustrate "production mode" by including it in the command line when running ISAS. The settings established in the last "interactive mode" session will be used. For example, at the operating system prompt, typing:

./ISASbasesNewCPU FILE=/home/Data/file1.fastq

will start ISAS, load the reference files, load the settings from the last ISAS session, load the database corresponding to the settings, and perform alignment on the file "/home/Data/file1.fastq". After ISAS finishes writing the results file, it will exit. This mode allows running ISAS in an automated fashion, e.g. from scripts.

4. FILES=x,y,min,max

This command is used to perform alignment on mated pairs of sequences. The first sequence in the file "x" and the first sequence in the file "Y" are considered a pair. Similarly, the second sequences in both files comprise the second pair, and so forth. "min" and "max" specify the minimum and maximum base pair distance allowed between the two sequences in a matched pair. For a pair of sequences to be found in the reference, three criteria have to be met:

a. Both sequences have to be found in the reference, with no more than AllowedMismatches (each).
b. Both sequences have to be found in the same chromosome.
c. The genomic distance between the two found sequences has to be at least "min" and no more than "max" (the last two parameters in the "FILES" command) apart.

The output file will have a ">" preceding the sequences of the first file and a "-" preceding the sequences of the second file, and also show the distance between each pair. The output file name will have "paired" at the end. The "FILES=" command can also be issued from the command line, just like the "FILE=" command (see 3 above).

5. MODE=x,y

This command is used to set the current search mode, where "x" is the Read Length and "y" is the maximum Allowed Mismatches. ISAS will save the current mode upon exiting, and re-load it the next time it is run, saving you the effort of typing this command every time, unless a change is desired. When INDEL (see 6 below) is disabled, the mismatches

are limited to substitutions. When INDEL is enabled, the mismatches can include any combination of mismatches, insertions, and deletions.

In any mode, searching for any sequence stops after a total of LIMIT matches are found for that sequence. LIMIT can be changed using the "LIMIT=" commands (see 10 below). Furthermore, any bases beyond ReadLength will be ignored.

6. INDEL=x

Use a value of 1 for "x" (i.e. INDEL=1) to allow indels (insertions and deletions) in addition to substitutions. This is often called "gapped" alignment, since an insertion in the sequence is equivalent to a gap in the reference, and a deletion in the sequence is equivalent to a gap in the sequence. If, for example, the maximum allowed mismatches is 4, then any combination (this is often called "edit ditance") of insertions, deletions, and substitutions, where the sum of all three types of mismatches is less than or equal to 4 is allowed as a match.

INDEL=0 disables indels, so only substitutions allowed. This is called "ungapped" alignment. Gapped alignment is computationally much more expensive than ungapped, and can be several times slower. It is slowest for larger values of maximum allowed mismatches.

INDEL=2 is a hybrid mode, where each sequence is first aligned without allowing indels, and if it is not found anywhere in the reference, it is re-aligned with indels. If the sequences are high quality, then most of them are found without indels, so this hybrid mode saves a substantial amount of run time as compared with INDEL=1 which is not recommended for this reason.

7. VERBOSE=x

Use a value of 1 for "x" to enable "Normal Output". In this state the output file will be as described in 3 above. Use a value of 2 for "x" for "Unique-Only", where the output file is as in "normal Output", except that only the sequences which were uniquely mapped are included. In other words, sequences which had no matches in the reference, as well as sequences with multiple matches, will be omitted from the output file. A value of 0 for "x" will switch to a "Non Verbose Output" state. In this state, the output file will only report the number of matches for each sequence, while the tag and the sequence itself will not be echoed to the output file. This can be useful for research purposes when comparing the effects of different settings. By comparing two "Non Verbose" output files produced from the same input file, but with different settings, specific sequences whose mapping characteristics have changed can easily be pin-pointed. VERBOSE=3 will write the output file in "SAM" format. More information about this format along with free software to perform various functions, including downstream analysis on this type of file, can be obtained at http://samtools.sourceforge.net. VERBOSE=4 will write SAM format, but only uniquely mapped sequences. When using "pileup" or SNP calling tools, make sure that the first line of chr1.fa is ">chr1", etc. You might have to change ">chrX" to ">chr23".

8. RANDOM=x

This command is used for creating files of sequences for testing and research. "x" specifies the number of sequences. Each sequence will first be copied from a random

location of the reference genome. Next, random mismatches will be added to the sequence, as defined by the current mode settings. The file name will begin with “Random”, show the Read Length, and total number of mismatches allowed, then “Ungapped” or “Gapped” depending on the current value of the INDEL setting (see 6 above). The output file name will end with “.txt”. The format will comply with the fastq format described in 3 above. Instead of an ID, the tag line will contain the “correct” alignment information for the sequence, i.e. the chromosome (if any) and genomic position in the reference from where it was originally copied. Approximately half of the sequences will be from the forward strand, and approximately half from the reverse strand. Reverse strand sequences will be denoted by a negative position.

9. RANDOMS=x

This command is similar to the “RANDOM” command (8 above) except that it will create a pair of output files. The second file will have a “2” appended to its name, before the “.txt” ending. Corresponding sequence in the two files will be on the same chromosome and strand, but separated by 1000 base pairs (as measured from their starting positions). Each file will contain “x” sequences, so there will be “x” sequence pairs, or 2 times “x” total sequences.

10. LIMIT=x

Some sequences could be repeated hundreds of thousands of times within the reference genome. Normally, it would be a waste of CPU time and disk space to search for and record all these locations. Furthermore, if we consider the alignment function is to uniquely identify the location of each sequence, then it would be most efficient to stop searching after two matches are found. ISAS still allows you to set LIMIT, which is the maximum number of matches searched for, to higher values if your application really requires this extra information. “LIMIT=x” will set the LIMIT to “x”. For each sequence, searching will stop if x matches are found. It is recommended to keep LIMIT as small as is practicable for your application. In other words, keep it at 2 unless your application requires finding more matches. Setting LIMIT to large values will incur a performance degradation.

11. REF=x,y

This function is very useful for research, and to investigate alignment behavior. ISAS will display the reference at chromosome “x” position “y”. If the reference has no chromosomes, use the value 1 for “x”. Reverse strands are denoted by negative values for “y”

12. CHR=x,y

Use this command to set the first chromosome to “x” and the last chromosome to “y”. If no chromosomes, use the value 1 for both “x” and “y”. For human, use “CHR=1,25” to include mitochondrial DNA, or “CHR=1,24” to exclude it. Normally this command only needs to be used once when setting up a new reference genome. ISAS will save this in the settings file when exiting, and re-load it when started the next time. You can also use this command to limit the search over a sub-set of all chromosomes, but the database for the smaller set will have to be created using MAKEBIN (see 14 below).

13. FASTA

Convert fasta files to reference files. This command only needs to be used once per new reference genome. You don't need it for the hg19 human reference genome as the reference files are included in the ISAS installation DVD for your convenience (See section III 2).

14. MAKEBIN

Create database files from reference files. This command only needs to be used once per new reference genome for each range of chromosomes. It can take 5 minutes or more, depending on your computer, for the full human reference. This database will be used each time alignment is performed.

15. HARDWARE

Display information about system hardware. ISAS will interrogate the hardware it is running on, and report the number of logical CPUs, amount of RAM, and the capabilities of the CPUs. This is a generally useful function, as not only is it easier than looking up your system documentation, it is also a more reliable source of current system information, as the system could have been changed from the last time it was documented (e.g. RAM added or removed, or CPUs upgraded).

```
Enter next command, or type "?" (and ENTER) for list of commands.
hardware

CPU: Intel(R) Xeon(R) CPU E5335  @ 2.00GHz (Stepping=7 Model=15 Family=6)
No HyperThreading. No PC. Has SSE3-SSE3 Extensions.
Has SSSE3-SSSE3 Extensions. No SSE4.1. No SSE4.2.
8 logical processors, 25,281,826,816 bytes RAM.

Enter next command, or type "?" (and ENTER) for list of commands.
```

16. FILTER=x

With FILTER=0, ISAS performs "lossless" alignment, also called "exhaustive" alignment. This means that no shortcuts are taken. If ISAS determines that there are X matches for a sequence, then there are exactly X locations in the reference which match the sequence within the maximum allowed mismatches. The only possible exception is when X reaches LIMIT. Filtering can optionally be used to further increase speed, by applying a shortcut to "problematic" sequences that are likely to be in repeat regions. Ten levels of filtering are available, from FILTER=1 (very slight filtering) to FILTER=10 (more aggressive filtering). Our experience (for human reference) has shown that using FILTER=3, alignment speed can be doubled by losing less than 0.1% of the sequences.

17. DATABASE=x

In its default, installed state, ISAS expects to be using the human genome reference known as “hg19”. This means that in the same directory where the ISAS executable (program) resides, there will be a sub-directory called “hg19”. This directory will hold the settings file, which stores your preferred settings (Mode, Verbosity, LIMIT, Filter, etc), as well as the reference database files, and a subdirectory called “reference”, which holds the reference files. The “DATABASE” command is used to switch to a different reference. The current reference directory is stored in a file called “ReferenceDirectory.txt” so when ISAS starts up, it will know which database directory to load. For example, to switch to the mouse reference mm9, type:

DATABASE=mm9

Note that in Linux, directory names are case sensitive. So the name of the directory cannot be “MM9”, it must be “mm9”. The structure and content of the directory “mm9” must be similar to the “hg19” directory, just with mouse data instead of human.

18. QUIT

Terminate the program. Just before terminating, the values of the following settings will be saved in a file called “settings.txt”:

FirstChromosome
LastChromosome
ReadLength
AllowedMismatches
INDEL
LIMIT
Verbose
Filter

This file will be in the current database directory. The current database directory name will be saved in the file “ReferenceDirectory.txt” in the executable directory. These files will be read back the next time the program is run.

19. EXIT

Same as “Quit” (16 above).

V. Performance Issues

1. BIOS Setting - Prefetch

IMPORTANT PERFORMANCE NOTE

Ignoring this will result in 40% slower alignment times.
On Xeon 53xx and 54xx systems: Turn off “Hardware Prefetch” and “Adjacent Cach Line Prefetch” in your BIOS settings.

Some computers are shipped with a default BIOS setting of “hardware prefetch” and “Adjacent Cache Line Prefetch” enabled. This causes the Xeon 53xx and 54xx CPUs to run ISAS (and many other programs which are memory intensive) 40% slower. If so, disable both of these in the BIOS settings utility program that you can invoke when you turn the power on. For example, in the PHOENIX BIOS utility program, these are in the “Advanced” tab, under “Advanced Processor Options”. We encountered this issue with many customers, so this is a common problem in popular brands like DELL.

2. Virtual Memory (Swapping)

If you don’t have the minimum amount of RAM (see section I 2) required, but still try to run a program, then the computer will try to substitute hard disk for RAM. This process is called “swapping” or “virtual memory”. Hard disk is orders of magnitude slower than RAM, and the overhead of swapping back and forth large parts of memory between RAM and hard disk for every small memory transaction can be overwhelming for the computer. The result is that the use of virtual memory causes programs to run **many times slower**, often causing the system to come to a virtual standstill. The reason virtual memory is used is that this situation is sometimes considered better than the alternative – which is a program crashing due to insufficient memory. However, in non-critical applications, it may be better to crash a program than to drag down the entire computer. In any case, it is not advisable to use ISAS with a reference genome which requires more RAM than you have. If you do, you must expect dramatic performance degradation.

Because ISAS uses a lot of RAM, Linux may mistakenly start swapping even though there is no need. Make sure to set Linux swappinness to zero (see page 5).

3. MultiTasking

IMPORTANT PERFORMANCE NOTE

When running alignment – do not run any other program that intensely uses the memory, hard disk, or CPUs.

CPUs can be “switched” back and forth between different programs with a switching overhead that can be reasonable if implemented correctly. ISAS is designed to utilize all the resources on your computer. If it has to “compete” with other programs, the performance of all of them will be greatly degraded. We recommend not running any other program except the Linux “top” program.

4. LIMIT (Maximum No. of Matches)

For each sequence, ISAS will keep searching the reference database until either there are no more matches (within the specifications of the current mode), or LIMIT matches have been found, whichever comes first. Normally, setting LIMIT=2 is all that is required to determine if a sequence can be uniquely mapped to the reference. Higher values for LIMIT will cause somewhat slower alignment times. Why would a LIMIT value higher than 2 be

used ? Multiple candidate locations for sequences that cannot be uniquely mapped might be of interest for some applications or research. For paired ("mate pair") runs, a higher limit (e.g. LIMIT=10) helps rescue pairs whose individual sequences fall in repeat regions.

VI. Imagenix Contact Information

The ISAS team at Imagenix has attempted to design and create a useful, easy to use product. Our hope is that it will improve your productivity, and may be one of many stepping stones to help you obtain results which will benefit all of mankind. However, we realize that there is *always* room for improvement, and we always welcome feedback from our customers.

If you have questions, comments, or suggestions of a technical nature, that you would like to share with us, we have found that the most effective means for this is generally by email, as it allows both sides time to think before replying.

When contacting us, please mention the version of ISAS you have. It is also helpful if you include a description of your hardware.

For technical support or feedback	technicalsupport@imagenix.com
For licensing information	sales@imagenix.com
General Information and updates	www.imagenix.com/genomics
Imagenix Technologies 171 Main Street #108 Los Altos, CA 94022	Tel. (650) 917-9998 Fax (650) 917-8765