Viral Integration Detection Power Calculator (VIpower)

User Manual

(Last edited: July 20, 2016)

Overview

Next generation sequencing (NGS) technologies are enabling rapid identification of integrated viral DNA sequences into the human genome. To assist investigators in designing high-powered NGS experiments aimed at detection of viral integrations, we created VIpower (viral integration detection power calculator). VIpower is a simulation-based power calculator, and consists of four modules: "Human sequence simulation", "Viral integration simulation", "In silico sequencing" and "Virus integration identification" (**Figure 1**).

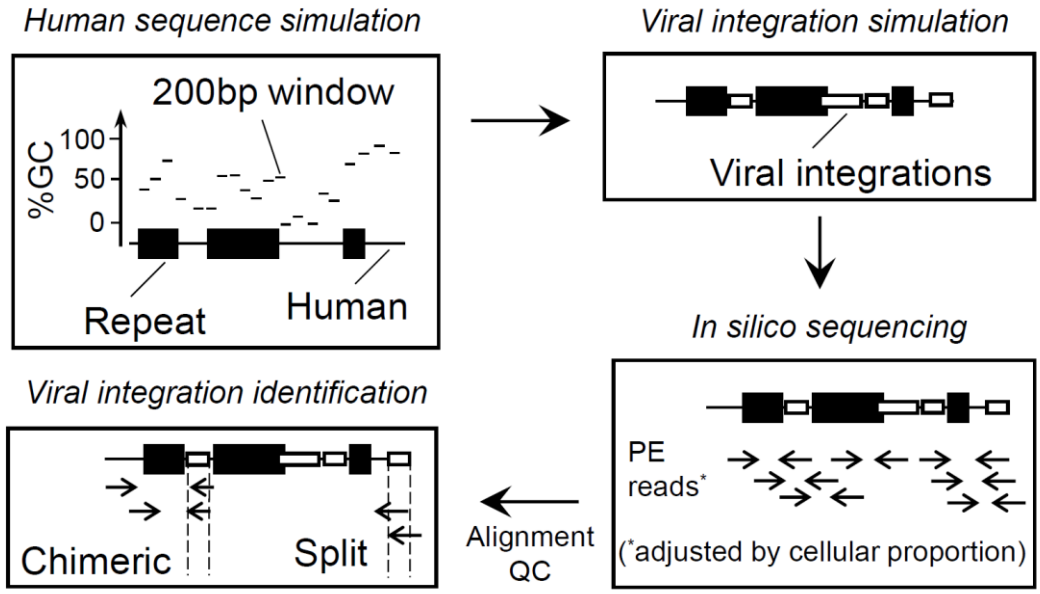


Figure 1: Overview of four modules underlying VIpower functionality

Intuitively, VIpower creates a "synthetic" human genome with GC and repeat region characteristics similar to those of the reference human genome (module 1), followed by assigning regions in that sequence as viral integrations according to a viral integration profile (module 2). The third module creates paired-end sequencing reads according to a read profile from a real sequencer (Illumina in this case). Lastly, chimeric and split reads are identified at each virus-human breakpoint (module 4), and finally the detection power is reported. Detection power is simply the proportion of viral integrations detected. Each module relies on empirical data (**Figure 2**).

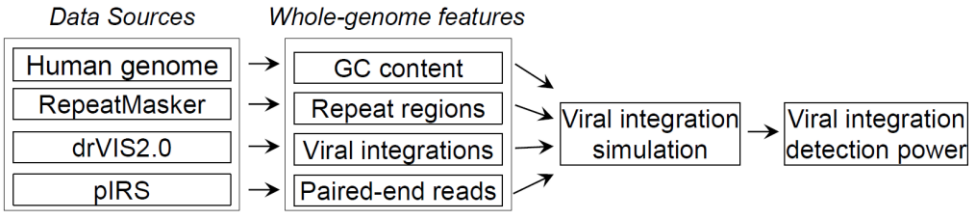
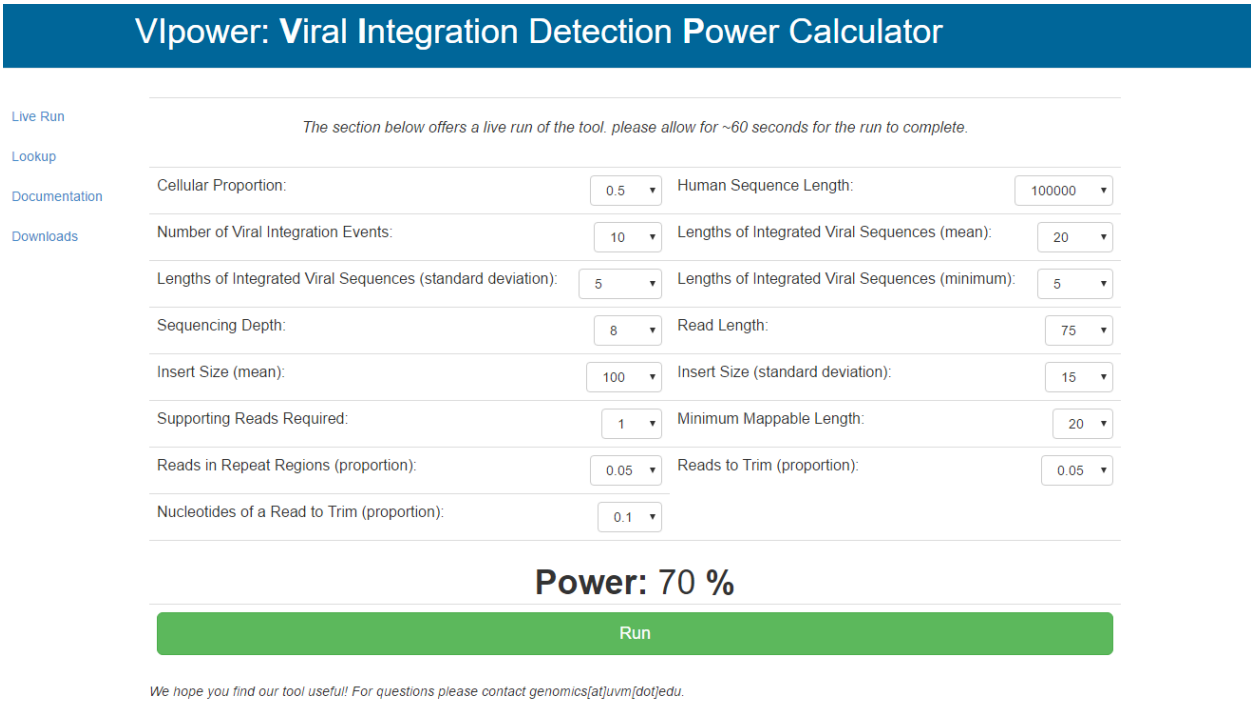


Figure 2: Empirical data sources used by VIpower.

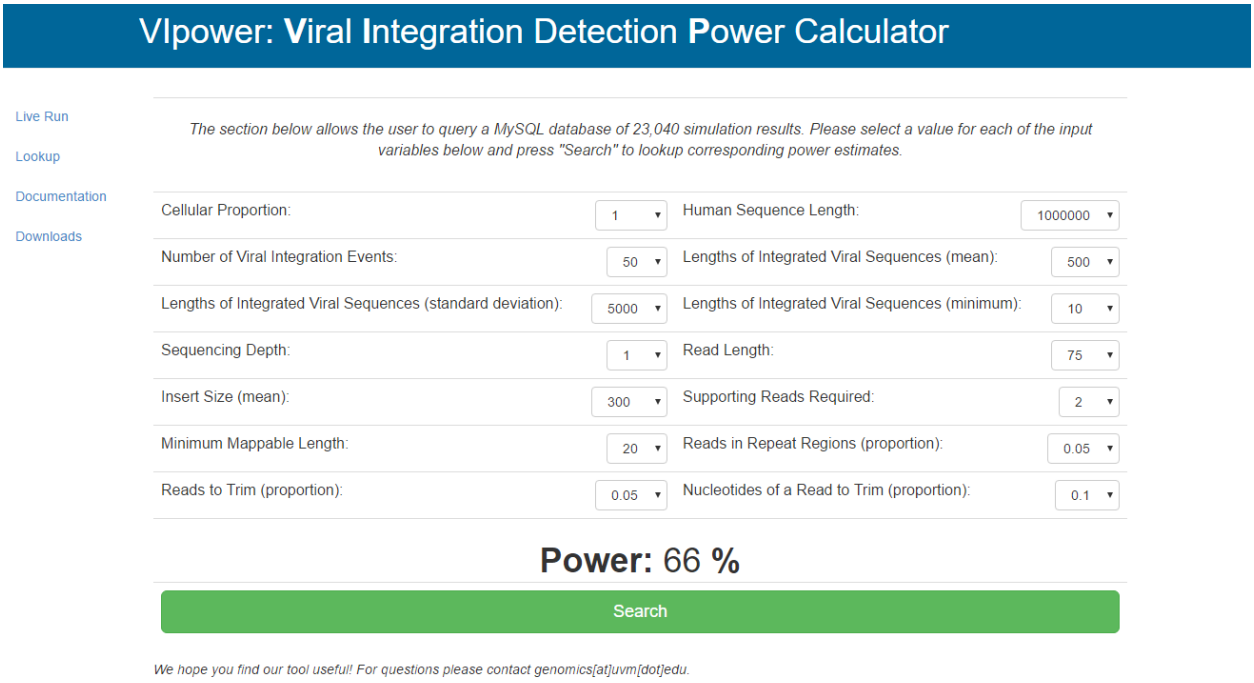
Running VIpower through the web

VIpower can be executed as a web-tool at http://www.uvm.edu/genomics/software/VIpower/live/. To keep computation time short (i.e., less than 60 seconds of wait time per run) we have assigned a discrete number of variables for each input. The available options encompass the range of reasonable values for each variables (except human sequence length, which can drastically increase computation time). The available values result in more than 10,000,000,000 unique variable combinations.



Querying precomputed power estimates

A total of 23,040 combinations of input parameters of VIpower were conducted and the table of results was saved into a MySQL database, which can be searched through the web at http://www.uvm.edu/genomics/software/VIpower.



Running VIpower locally

The user has two options for running VIpower locally: (1) standard and (2) advanced option.

(1) Standard option

i. Download RData file:

```
wget -c
http://www.uvm.edu/genomics/software/VIpower/downloads/Stan
dard_simulation_workspace.RData
```

ii. Download simulation script:

```
wget -c
http://www.uvm.edu/genomics/software/VIpower/downloads/vipc
.r
```

iii. Run VIpower with user-supplied variables

```
Rscript --vanilla vipc.r [arg1] [arg2] [arg3] [arg4] [arg5]
[arg6] [arg7] [arg8] [arg9] [arg10] [arg11] [arg12] [arg13]
[arg14] [arg15] [arg16]
```

, where the order of each argument [arg] is as follows:

[arg1] - Human Sequence Length (including total length of viral insertions)

[arg2] - Total number of viral integrations

[arg3] - Average length of viral integrations

[arg4] - Standard deviation (SD) of viral integration length

[arg5] - Minimum viral integration length

[arg6] - Sequencing depth (fold)

[arg7] - Read length (bp)

[arg8] - Average insert size (bp)

[arg9] - SD of insert size (bp)

[arg10] - Number of supporting PE reads per viral integration (chimeric + split)

[arg11] - Minimum mappable length (bp)

[arg12] - Proportion of PE reads falling completely inside repeat regions

[arg13] - Proportion of a read length that is trimmed

[arg14] - Proportion of reads that will be trimmed

[arg15] - Cellular proportion

[arg16] - Random seed

Additionally, here is an explanation for each argument.

Table 1: List of command line parameters of VIpower.

Parameter	Description
human_virus_length	Total sequence length, including integrated viral sequence (bp)
vip_nr	Number of viral integration events
vip_len_mean	Average length of viral integrations (bp)
vip_len_sd	Standard deviation of lengths of viral integrations (bp)
vip_len_min	Minimum length of viral integrations (bp)
seq_depth	Sequencing depth (fold or X)
read_length	Read length (bp)
read_insert_mean	Average of insert size (bp)
read_insert_sd	Standard deviation of insert size (bp)
read_perVIP	Required minimum number of supporting (chimeric and split) reads
min_read_length	Minimum read length uniquely mapped to either human or viral reference
read_repeat_freq	Maximum proportion of reads completely mapped inside repeat regions
read_trim_prop	Proportion of total reads that are trimmed
read_trim	Proportion of read length that is trimmed
cell_prop	Proportion of cells with viral integrations (e.g., somatic, <1 and germline, 1)
seed_value	Simulation seed (for reproducible results)
ltr_array	Repeat sequence distribution (~5.2 million repeat regions from RepeatMasker)
gc_array	GC content distribution specific to the human genome

(2) Advanced option

This option is recommended for users who want to change the reference files. For instance, if the PE reads' profile is to be modelled after a different sequencer from Illumina, or if the viral integration profile is to be modelled after a different virus from HBV, then the reference files may by modified by the user.

i. Download all R scripts and supporting reference files.

```
wget -c
http://www.uvm.edu/genomics/software/vip/downloads/vipc_com
plete.zip
```

ii. Unzip the package

```
unzip vipc_complete.zip
```

iii. To create a new virus integration profile: replace content of `HBV_GC_profile_array.txt` file with the new virus profile. Makes sure the column headers remain the same.

iv. To create a new profile based on distance from repeats to virus integration breakpoint: replace content of `ltr_vip_distance.txt` with the new values.

v. To change sequencer-specific PE-read whole-genome distribution by GC content, replace content of `pIRS_GC_coverage_VIPS_200bpWindow.txt` with new values.

vi. Run `Rscript --vanilla SIM_BUNDLE.r [arg1] [arg2] [arg3] [arg4] [arg5] [arg6] [arg7] [arg8] [arg9] [arg10] [arg11] [arg12] [arg13] [arg14] [arg15] [arg16]`

, where each of the arguments follows the same order as those in "Running VIpower locally".

Notes: Both standard and advanced options require that you have `R` and `Rscript` installed in your system. Setting the human sequence length to ≥100M bp may result in long wait times. However, we have shown that length of the human sequence does not impact detection power.

Frequently asked questions

Q1. What is VIpower?

VIpower (Viral Integration Power Calculator) is a simulation based power calculator for viral integration detection using NGS. It simulates characteristics of a human genome with viral integrations by sampling from distributions of empirical genomic data. It then uses empirical NGS sequencing information to conduct "*in silico alignment*" followed by viral integration detection, measured by presence of supporting reads (chimeric or split).

Finally, VIpower outputs a viral integration detection power estimate for a given NGS experiment.

Q2. Do I need to provide any personal information to use VIpower?

VIpower is publically available for (non-profit) use by anyone with access to the internet.

Q3. How do I use VIpower to calculate viral integration detection power?
There are at least two ways to run VIpower: through the web and on a local computer. We recommend using the web-version of the tool, as it should satisfy most users' needs. See sections 3 and 4 above for more information.

Q4. How do I estimate standard deviation of power estimates?
Power estimates from a single run of VIpower will generate a power estimate with no standard deviation or error information. To get a confidence interval for a given NGS experiment, we recommend running VIpower multiple times by changing only the random seed variable each time. The resulting power estimates may then be used for a confidence interval estimate.

Q5. How do I measure detection power for a single viral integration event?
After extensive simulations, we found that the number of viral integrations did not affect power calculations (as long as viral integration density does not change significantly). Thus, the power estimates for 1 viral integration or 10 or 100 will not be significantly different.

Q6. Who will use VIpower?
We hope VIpower will become a useful tool to many members of the scientific community from investigators applying for funding on projects that rely on viral integration detection, to investigators interested in detecting viral integration events from cancer biopsy samples or bioinformaticians designing novel tools for viral integration detection. We hope VIpower helps design powerful NGS experiments for detection of

Q7. How do I cite VIpower?
VIpower manuscript is currently under submission. We will make the reference available as soon as the manuscript is in press.