Variome Average

(变异组指数 命名参考：Dow Jones Industrial Average )

# For a haploid genome,

suppose:

CDS region ratio a ∈ (0,1)

CDS region effect m

Average codon-shift value s

Strandard regulation region ratio b ∈ (0,1)

Regulation region effect n

Other region effect o

Genome Size g

Het SNP weight Whs

Het Indel weight Whi

SV Confidence v ∈ (0,1)

Complex SV weight c ∈ (0,1)

Average CDS length d

There should be:

.

Also,

CDS region length(CRL), CDS count(CC), Gene region length(GRL) and Gene count(GC) can be directly counted from the annotation file, such as GFF files.

Since a ordinary human gene come with a regulation region of less than 12.45k (according to Affymetrix GeneChip® Human Tiling 2.0R Array Set), and most cis-elements are separated …, we choose an average value of 750 bp for the cis-elements. Also, each CDS, which comes with an exon, will require 4 bases for the splice signal at both ends.

Thus, . .

We can also calculate the actual amount of Variome in CDS region for a annotated genome.

We set those parameters:

m=1

s=1.1

n=0.9

o=0.001 these 4 effects cannot be explained..

Whs=0.5 Het SNPs are half in the amount

Whi=0.6 It is possible for a Het Indel be two different indels, thus bigger than 0.5

v=0.5 See below.

c=0.65 Different SV may share the same region.

# For SNPes

∀ xi∈SNP, the probability for x to be in CDS region is a, in regulation region is b,

thus the impact from xi is ;

HET SNPs are count as Whs each. And each SNP count is divided by its copy number.

So, for all x SNPes, 

The average value for SNPes will be: .

# For Indels

∀ indel with a length of l, thus the average length in CDS region will be , in regulation region will be , and in other region will be 

In every CDS, there is a probability of  for codon-shifting if every indel in CDS region is shorter than the CDS (), which is obviously true.

Thus the count of codon-shifting is 

∵The count of CDS region is 

∴ The effect of codon-shifting is 

Thus the impact is .

HET Indels are count as  each.

The average value for Indels will be: .

# For SVs

SVs can be of complex types. If more than 1 type appeared, each type is calculated separately with a weight of c.

SV Confidence is set because SVs from soapSV cannot be specify to certain base pairs, the actually position is not aviliable. Without further information, we just set v to 0.5.

Transposons will be calculated as half its length.

∀ SV with a length of l, thus the average length in CDS region will be  for non-transposons and  for transposons.

Like indel,

The average value for SVs will be:  for non-transposons and  for transposons.

Complex SVs are calculated as above and then sum up by a weight of c.

# The average

To make a average of SNPes, Indels and SVs, we must add them together. Since SV is for a large fragment while either SNP or Indel is for a small number of bases, these two 者相差sereval量级。Thus, we apply a weight that 对SV的权重比其他的少200，以平衡量级。

(由于SV权重与v地位相同，可按需调整二者。)

So,

SNP weight: 1000

Indel weight 1000

SV weight 5

Thus the total average is: 

（原则就是最终3个数，SNP>Indel>SV， SNP>1。）

对单倍体基因组，设：

CDS比例: a ∈ (0,1)

CDS区影响权重pre bp: m

平均移码影响值（pre CDS）: s

标准化调控区比例（含启动子、增强子加权值）： b ∈ (0,1)

调控区影响权重pre bp: n

基因组大小： g

sv不确定度： v ∈ (0,1) $SV\_Confidence

复合sv拆分系数： c ∈ (0,1)

平均CDS长度 d

则有：

a+b<1

CDS比例=CDS总长/基因组序列总长。 基因组用原始的。

标准化调控区比例=基因数量\*750+CDS数量\*4. 基因数与CDS数源于GFF。一般找启动子是上游2k以内，也有1.5k的，这里就750了。剪切位点是GT/AC，虽然CDS区只占2个，但一起算比较方便。

SNP：

∀ 位点xi，其在CDS区的概率为a，在调控区的概率为b，

故产生的影响为 a∙m+b∙n

对x个SNP，∑，有x(am+bn);

按基因组大小归一化，得：x(am+bn)/g.

Indel:

∀ 长度l，其在CDS区的平均长度为al，在调控区的平均长度为bl，

∵每个位于CDS区内的片段有2/3的概率造成移码（al<d）

∴移码次数为2al/3d.

而CDS区有ga/d个，

故移码影响为 (2al/3d \* s)/(ga/d)=2ls/3g .

故，总影响：2ls/3g + al/3g + bln/g。

对 总长为 l 的indel，∑，

得：l(2s + am + 3bn)/3g。

SV (由于有复合的，只能每个单独算):

sv不确定度是指文件中报道长为l的区域为包含某类sv的区域，但实际上其中生物学sv的长度只有vl < l。

对复合类型的sv，分别计算，再乘以c加权相加。根据复合数量N设置c，使1 < Nc < N就可以了。

∀一个SV，覆盖CDS区部分长为vla，包括 vla/d 个CDS。

转座：

转座元件长lv/2

倒位区、重复区、Insertion/Deletion区长 lv

将长度代入SV的公式，得：

转座： lv(2s + am + 3bn)/6g。

倒位、重复、Insertion/Deletion： lv(2s + a + 3bn)/3g。

复合的SV，累加后乘以c。

New：

$U\_m=1; # CDS区影响权重pre bp

$U\_s=1.1; # 平均移码影响值（pre CDS）

$U\_n=0.9; # 调控区影响权重pre bp

$SNP\_HET\_Ratio=1.6; # 杂合SNP系数

$Indel\_HET\_Ratio=0.9; # 杂合Indel系数

$SV\_Mixture\_Ratio=0.75; # 杂合Indel系数

%SV\_Weights=( # 复合sv拆分系数

Transposion => 0.5,

); # 转座的权重算1/2

$SV\_Confidence=0.25; # SV可靠度 / sv不确定度:v

$SNP\_r=1000;

$Indel\_r=1000;

$SV\_r=10;

$U\_a=$CDSLen/$GeneLen; # CDS比例;

$U\_b=(750\*$GeneCount+4\*$CDSCount)/$GeneLen; # 标准化调控区比例（含启动子、增强子加权值）

$U\_SNP=$SNP\_r\*$CountSNP\*($U\_a\*$U\_m+$U\_b\*$U\_n)/$GeneLen;

$U\_Indel=$Indel\_r\*$CountIndel\*(2\*$U\_s + $U\_a\*$U\_m + 3\*$U\_b\*$U\_n)/(3\*$GeneLen);

$U\_SV=$SV\_r\*$CountSVA\*(2\*$U\_s + $U\_a\*$U\_m + 3\*$U\_b\*$U\_n)/(3\*$GeneLen);

$U\_mark=$U\_SNP+$U\_Indel+$U\_SV;

CDS比例 $U\_a=CDS总长/基因组序列总长。 基因组用原始的。

标准化调控区比例 $U\_b=基因数量\*750+CDS数量\*4. 基因数与CDS数源于GFF。一般找启动子是上游2k以内，也有1.5k的，这里就750了。剪切位点是GT/AC，虽然CDS区只占2个，但一起算比较方便。

3个权重：

$SNP\_r=1000;

$Indel\_r=1000;

$SV\_r=10;

SV的低100，1000是为了调到1～10量级。

SNP得分：

$U\_SNP=权重\* SNP个数\*(CDS比例\*CDS区影响权重+标准化调控区比例\*调控区影响权重)/基因组大小;

$U\_Indel=权重\* Indel标准长度\*(2\*平均移码影响值 + CDS比例\*CDS区影响权重+ 3\*标准化调控区比例\*调控区影响权重)/(3\*基因组大小);

$U\_SV=权重\* SV标准长度\*(2\*平均移码影响值 + CDS比例\*CDS区影响权重+ 3\*标准化调控区比例\*调控区影响权重)/(3\*基因组大小);

Indel标准长度，杂合的按$Indel\_HET\_Ratio算

SV标准长度：复合的各自乘以$SV\_Mixture\_Ratio后累计。转座的再除以2。最后总和乘以$SV\_Confidence。