Recurrence analysis considering sequence composition and variable mutation rates

Inigo Martincorena

ICGC pancancer phone call

17th Nov 2014

Factors to consider

1. Length of the element

Factors to consider

2. Trinucleotide mutation rates and sequence composition

E.g.: C>T at CpGs tend to have a rate per bp ~40x higher than other substitutions

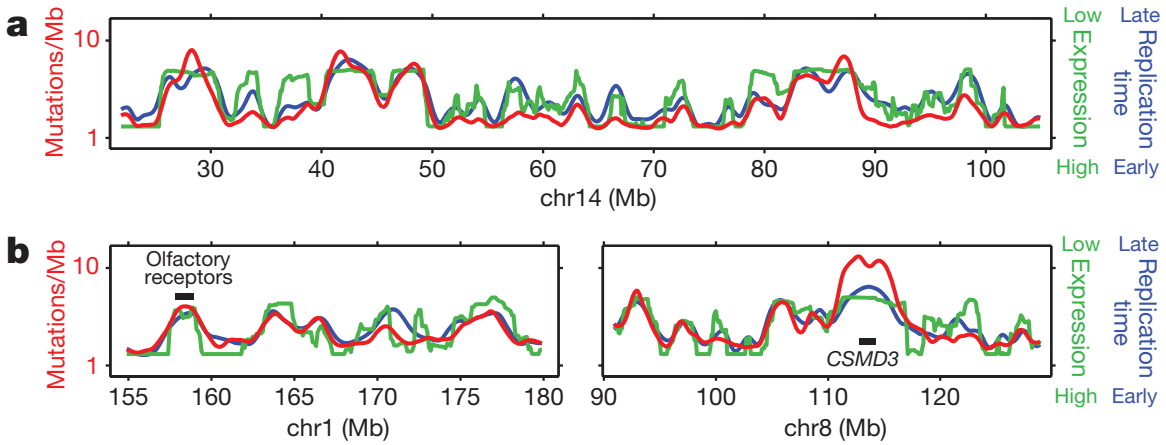
Random mutations:

Element 1

Element 2 (e.g. TFBS, phosphorylation site…) could be significant with uniform models

Factors to consider

3. Regional variation of the mutation rate



Possible solutions:

- Use local density of mutations (e.g. silent sites, 10kb around…)
- Use covariates (*e.g*. expression, repl time…)

But uncertainty in the estimates needs to be included in the calculation of pvalues.

607 pilot (gene analysis only)

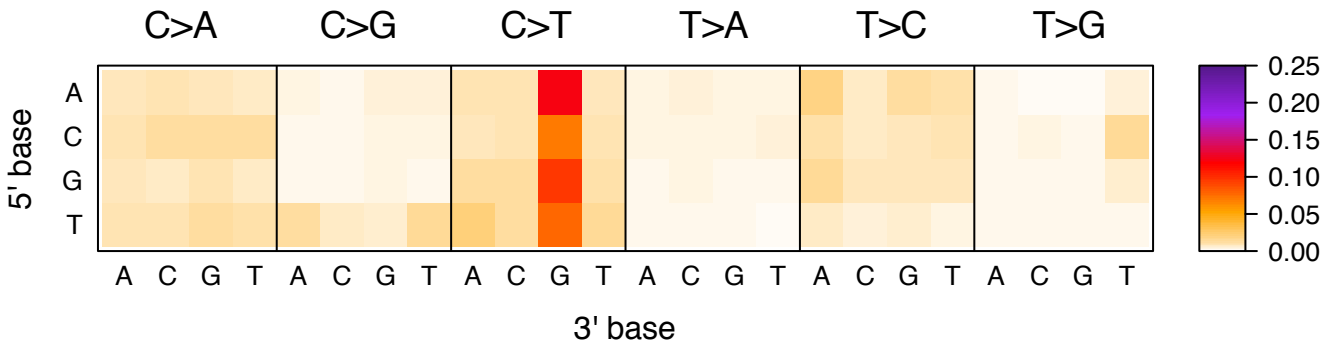
Simulation of random data to test the specificity of the models

I have simulated a random version of the 607 pilot while retaining:

- Same trinucleotide rates
- Same regional variation of mutation rates (per Mb)
- Same distribution of mutations across samples and tissue types

Not as challenging as true data. But significant hits with low q-values would evidence that the background models are inadequate.

https://www.dropbox.com/s/ci9vhbr5xe3g2wc/simulated_607_dataset.tgz

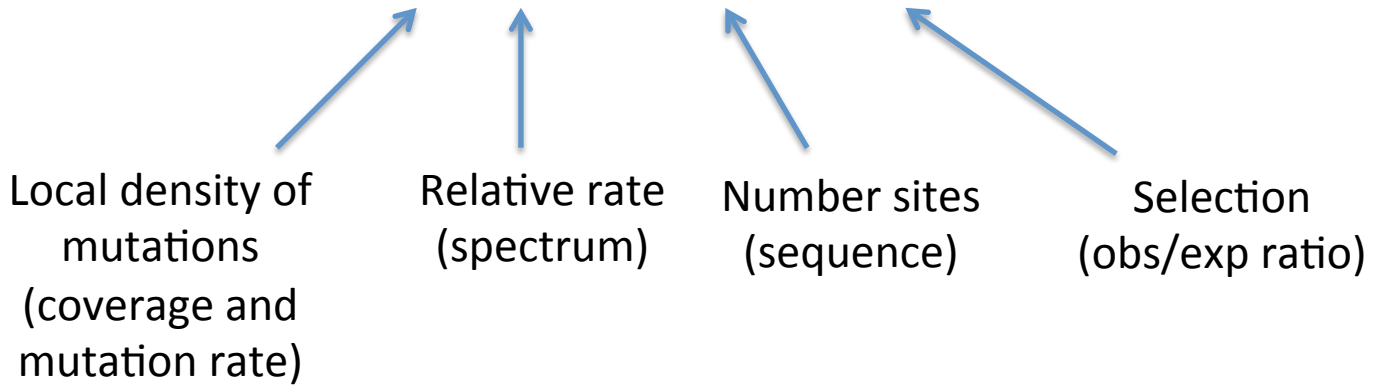


Our approach on genes

e.g. frequency of synonymous and missense A>C mutations

$$\lambda_{syn,AtoC} = (t)*(AtoC)*(L_{syn,AtoC})$$

$$\lambda_{mis,AtoC} = (t)*(AtoC)*(L_{mis,AtoC})*(wMIS)$$



Greenman *et al*. Genetics, 2006

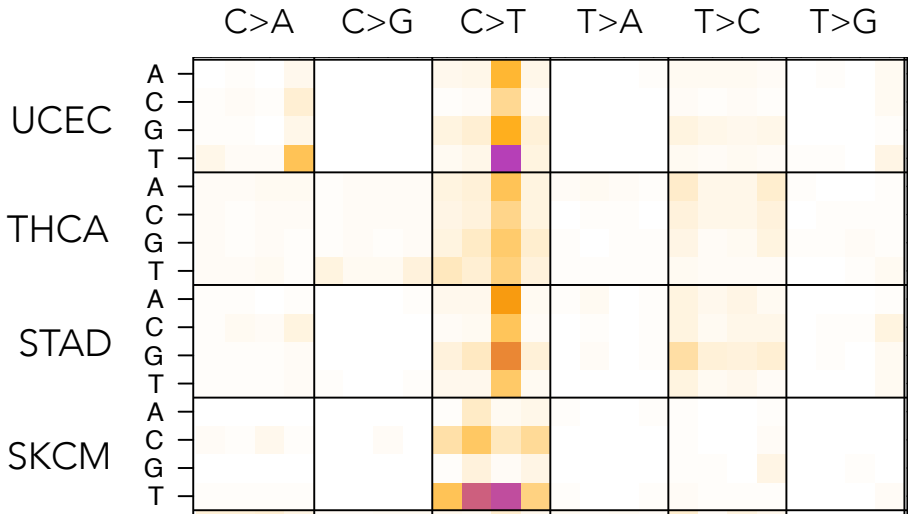
Adding context-dependence

Context effects easily incorporated as multiplicative rates

$\lambda_{mis,CtoT} = (t)*(CtoT)*(L_{mis,CtoT})*(wMIS)$ *12 rates model*

$\lambda_{mis,CtoT} = (t)*(CtoT)*(L_{mis,CtoT})*(wMIS)*$**(CpGctx)***(TpCctx)** *16 rates model*

$\lambda_{mis,CtoT} = (t)*(ACGtoATG)*(L_{mis,ACGtoATG})$ *192 rates model*



Local mutation rates

Number of coding mutations

1,000 — 10,000 — 100,000 — 1,000,000

Pan-cancer

False positives

Low sensitivity

Model 1: **bg = t**

Assumes no variation between genes.
Gives hundreds of FP in large datasets.

Model 2: **bg = Pois(t|n_syn)**

Very conservative.
Clean results for large datasets.
It does not exploit info from other genes.

Local and global model

Exploiting covariates: negative binomial regression (t*trinuc_rates*L)

- Trinucleotide rates and sequence composition included as the offset (fixed factors)
- Negative binomial: mutations are Poisson, rates among genes Gamma distributed
- Adding covariates reduces the uncertainty (dispersion) of the background -> higher power

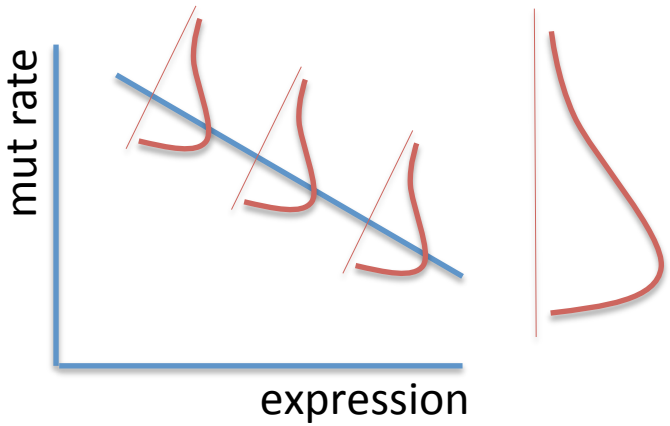
Covariates are fitted to each dataset, not assumed

bg = Gamma(t|covariates)

Negative binomial regression

glm.nb(obs ~ offset(expected) + expr + reptime …)

	Obs_syn	Exp_syn	Expression	Repl time
Gene A	10	8.53	20012	4.35
Gene B	38	30.76	1782	2.67
Gene C	1	2.45	59927	-8.6
…	…	…	…	…
Gene Z	290	263	306	4.56



Local mutation rates

Number of coding mutations

1,000 — 10,000 — 100,000 — 1,000,000

Pan-cancer

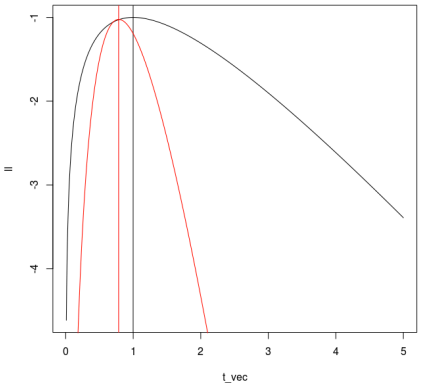
bg = Pois(t|n_syn) * Gamma(t|covariates)

expression

chromatin

rep time

any other



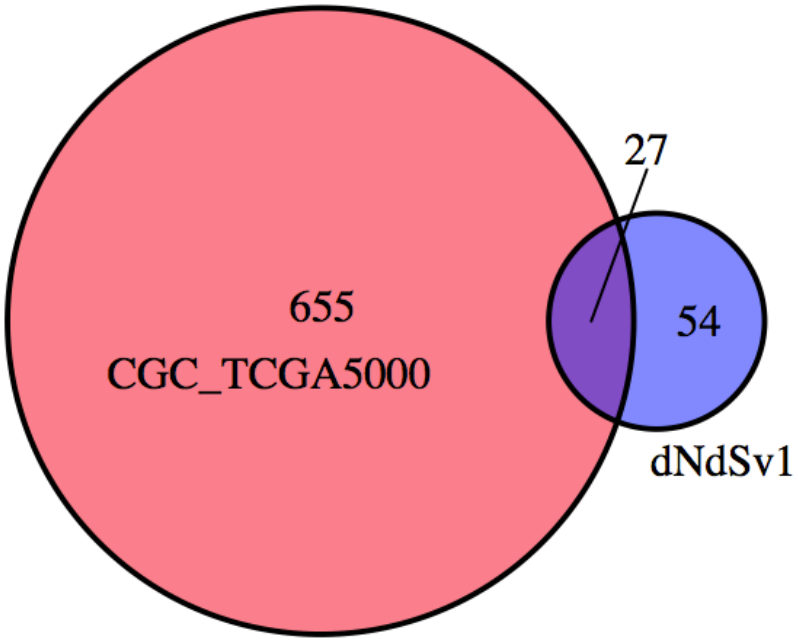
Model 3

bg = Pois(t|n_syn) * Gamma(t|cov)

bg = Pois(t|n_syn) * Gamma(t|cov)

607 pilot (gene analysis only)

A) Only correcting for sequence composition



Overlap = 33.3%

Dubious novel cancer genes as the list includes TTN and other large genes as well as interesting candidates.

Random dataset:

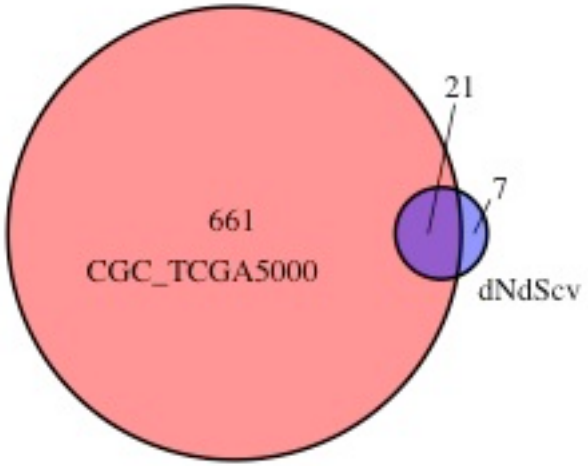
9 false positives

Lowest qval = 4e-9

Obviously the background model is not realistic

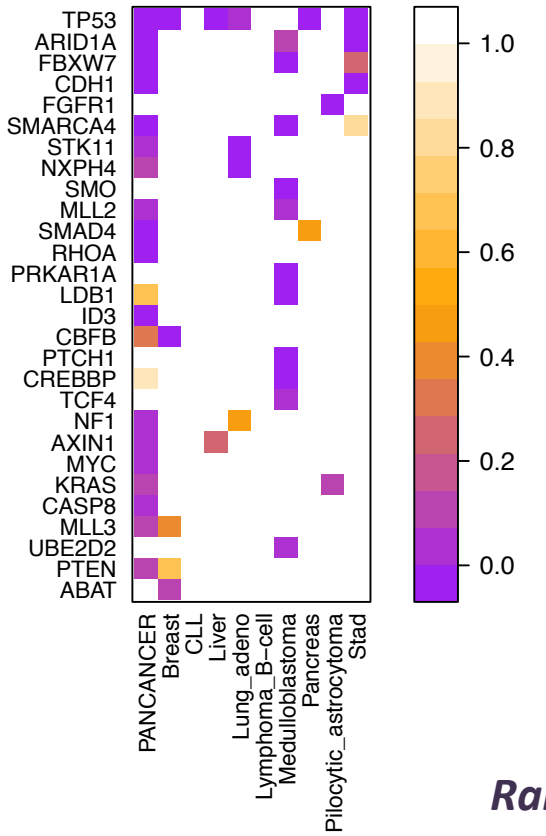
607 pilot (gene analysis only)

B) Accounting for sequence composition and variable mutation rates across genes



Overlap = 75%

Good novel cancer genes: At least 5 of the 7 new candidates are known to be involved in cancer (PTCH1, ID3, LDB1, TCF4 and ABAT).



Random dataset shows very good behaviour

0 false positives

Lowest qval = 0.86

607 pilot (gene analysis only)

Some conclusions and ideas for noncoding sites:

- Need to account for extreme trinucleotide rates: especially for hotspot analysis and recurrence in small motifs (e.g. TFBS, phosphosites, functional bias…)

- Need to account for variable mutation rates
  - *Local estimates and/or covariates*
  - *And explicitly consider the uncertainty that remains in the model*

- Increasing power within this framework:
  - *Site recurrence or local hotspots (same model applies to a single site)*
  - *Classify sites by importance: nonsense vs missense, polyphen, TFBS within promoters…*

- FDR correction has to be decided *a-priori* (*e.g.* correct for all genes/sites even if p-values were calculated only for those than had mutations)