Coalescence modeling of cancer sequencing data

Andrzej Polanski
Silesian University of Technology, Gliwice, Poland
15 September, 2016
PLAN

• Coalescent model
• Motivation for use in cancer sequencing data
• Methods
• Results
• Conclusions
What is coalescent model?

Coalescent is a mathematical tool for modeling phylogenetic trees of samples drawn from evolving populations. It has been used for statistical inference on genetic parameters and structures of evolving populations.
What is coalescent model?


Wright-Fisher model of reproduction

Individuals = fragments of DNA

Alleles: $A_1$: ○ $A_2$: ●

Reproduction = DNA replication = sampling with replacement

Hypotheses:

- Discrete, non-overlapping generations
- Haploid individuals or diploid population with panmixia = random mating
- Constant population size
- No selection (all individuals are equally fit)
- No recombination
- No substructure (no geographical structure)
Wright-Fisher model: reproduction + mutation

Mutation introduces genetic variability to the evolution process

- Poisson process with intensity \( \mu \) measured per locus (per site) per generation. Spatial characterization of places and effects caused, further specifies a mutation model. Most often applied:
  - Infinite sites model, where it is assumed that each mutation takes place at a DNA site that never mutated before;
Coalescent approach

One looks at the past of an $n$-sample of sequences taken at present. Possible events that happen in the past are coalescences leading to common ancestors of sequences, and mutations along branches of ancestral tree.
Coalescent approach

\[ G_1 \]

\[ G_2 \]

\[ \ldots \]

\[ G_n \]

1 2 3 4 5

\[ 1 \quad 2 \quad 3 \quad 4 \quad 5 \]

\[ t_2 \]

\[ t_3 \]

\[ t_4 \]

\[ t_5 \]

\[ s_2 \]

\[ s_3 \]

\[ s_4 \]

\[ s_5 \]
Coalescent versus other approaches

Coalescent

- Analytical results
- Coalescent simulations

Branching processes

Forward random simulations
Coalescent

- Changing population size
- Selection
- Substructure
Coalescence, topology, metrics

Mutations

Sequences

Metrics: Continuous (diffusion) approximation
Type of mutation.  
Mutation no 4 is of type (5,2)
Types of mutations (alleles)
SNV allelic frequencies
Allelic frequency spectra

Mutation \rightarrow \text{Mutation type } = (5,2)


Coalescence modeling for cancer genomics data

The Cancer Genome Atlas, TCGA database

GDC database
TCGA, BAM files

.BAM .VCF mutation column
Coalescent trees for somatic mutations

N = 22 reads (size of the coalescence tree), b=11
Methods
Deriving SNV allelic frequencies requires mathematical modeling of both topology and (full) metrics of coalescence trees.
Probability that a mutation is of type \( b \)

**TREE TOPOLOGY**

Depends on metrics of the tree:

\[
q_b = \frac{\sum_{k=2}^{n} p_k^n(b)kE(S_k)}{\sum_{k=2}^{n} kE(S_k)}
\]

Probability that a mutation at the level where there are \( k \) ancestors will grow to \( b \) copies at the bottom of the tree

\[
p_k^n(b) = \binom{n-b-1}{k-2} \binom{n-1}{k-1}
\]

TREE metrics

\[ p(t_2, \ldots, t_n) = \prod_{k=2}^{n} \frac{k}{2} \frac{N(t_k)}{N(t_{k+1})} \exp \left( - \int_{t_k}^{t_{k+1}} \frac{k}{2} \frac{N(\sigma)}{N(t_{k+1})} d\sigma \right) \]

\( N(t) \) – history of population size


• Polanski A., Szczesna A., Garbulowski M., Kimmel M., Coalescence computations for large samples drawn from populations of time-varying sizes, submitted.

\[
\mathfrak{S}\{p(t)\} = P(s) = \int_0^\infty p(t) \exp\left(-s \int_0^t \frac{d\sigma}{N(\sigma)}\right) dt
\]

\[
\mathfrak{S}^{-1}\{P(s)\} = p(t) = \frac{1}{2\pi j} \frac{1}{N(t)} \int_{c-j\infty}^{c+j\infty} P(s) \exp\left(\int_0^t \frac{d\sigma}{N(\sigma)}\right) ds
\]
Growing versus constant – SNV allelic frequencies
Allelic frequencies, constant scenario:

\[ N(t) = N_0 \]

Past

Present

Population size

Reciprocals of successive integers

Stationary

Does not depend on \( N_0 \)
Allelic frequencies, exponential scenario:

\[ N(t) = N_0 e^{-rt} \]

\[ \rho = rN_0 \]
Results
TCGA Data processing pipeline

Programs, platforms

- e.g. Illumina
- bwa
- samtools

Data formats

- .FASTQ
- .SAM
- .BAM
- .VCF

Sequence alignment map

Binary alignment map

Variant calling format

Data compression

Reference genome alignment

Seqencing data

Mutations calling

- Mutect 1
- Mutect 2
- Muse
- Somatic Sniper
- Varscan 2
Somatic mutations (somatic intra-clonal mutations)

- Mutations
  - Germline
  - Somatic
    - Somatic in normal cells
    - Somatic in cancer cells
Somatic cancer mutations, Clonal growth model, subclones

- Clonal driver mutation
- Sub-clonal driver mutation
- Clonal passenger mutations
- Sub-clonal passenger mutations
- Intra-clonal mutations
- Intra-sub-clonal mutations

Normal tissue

Clone

Subclone

Time
Somatic mutations, Mutect 1

.BAM file cancer
.BAM file normal

Mutect

List of detected somatic mutations in cancer cells
List of all detected mutations

Somatic mutations, Mutect 1

TCGA-02-0003 GBM WXS data

Mutect

List of all detected mutations

List of all detected somatic mutations in cancer cells

606 145

247
Clonal growth model

histograms of VAFs

Intra – clonal mutations

Clonal mutations

sub-clone

cloned

Intra-clonal mutations

Normal tissue

clone

subclone

No of som. muts: 297  KS = 3  SurvTim = 14.72
Intra–clonal mutations
Distributions (histograms) of allelic types in TCGA datasets
Analysis (TCGA-02-0047) (GBM patient) Mutect 1

• Filter all Mutect 1 mutations such that:
  - coverage is > 20
  - mutation is unique to cancer cells
Analysis (TCGA-02-0047) GBM, Mutect 1
Problem no 1: shortage of alleles of type $b=1$

Alleles of type $b=1$ are possibly excessively removed by Mutect 1 filtering algorithm

Remedy: condition on $b \geq 2$
Problem no 2: contamination by normal cells

- Contamination by normal cells introduces spurious reads, which are aligned to the reference genome
- Can require additional modeling (estimating tumor purity)
Problem no 3: excess of rare alleles (variants)

In all 50 GBM WXS datasets it was not possible to fit exponential growth model to sequencing data.
Problem no: 3, Explanation no: 1
Recent population explosion

Unlikely, many papers estimate exponential or sub-exponential tumor growth.
Problem no: 3, Explanation no: 2

Clonality


All clones must be very small. Not seen in data.
Problem no: 3, Explanation no: 3

Weak selection. Cumulative effect of weakly deleterious passenger mutations
Tug of war between driver and passenger mutations
Cells with larger number of (intra-clonal) passenger mutations are under selection pressure.

Cells with their intra–clonal passenger mutations:
Experiment: ms program was used to generate 1000 cells under constant population size scenario with $\theta=100$, the selection with threshold 70 was applied.
More experiments, GDC database

• Mutect 2  
• Muse  
• Somatic Sniper  
• Varscan 2  

• VEP (Varian Effect Predictor)

• Do we observe differences between frequency spectra in:
  - different cancer types  
  - different patients  
  - different DNA sites ?
**Aggregate statistics**

- **Lung cancer Intron variant**
  - Coverage: 789

- **Lung cancer Missense variant**
  - Coverage: 151
Conclusions

• Possibility of using coalescent theory to model cancer genomics data was mentioned by several authors (N. Boerwinckel et al. 2014, R. Durrett 2013).
• Many challenges, due to specificity of cancer genomics data
• In particular weak selection hypothesis may be necessary
• If the tug-off-war hypothesis is true, coalescence modeling with weak selection of passenger mutations model together with using .VCF data may contribute with estimations of parameters of weak selection process
• Weak selection hypothesis seems to be supported also by other arguments
• Perspective of correlating differences between patients with phenotypes
Thank you for your attention!