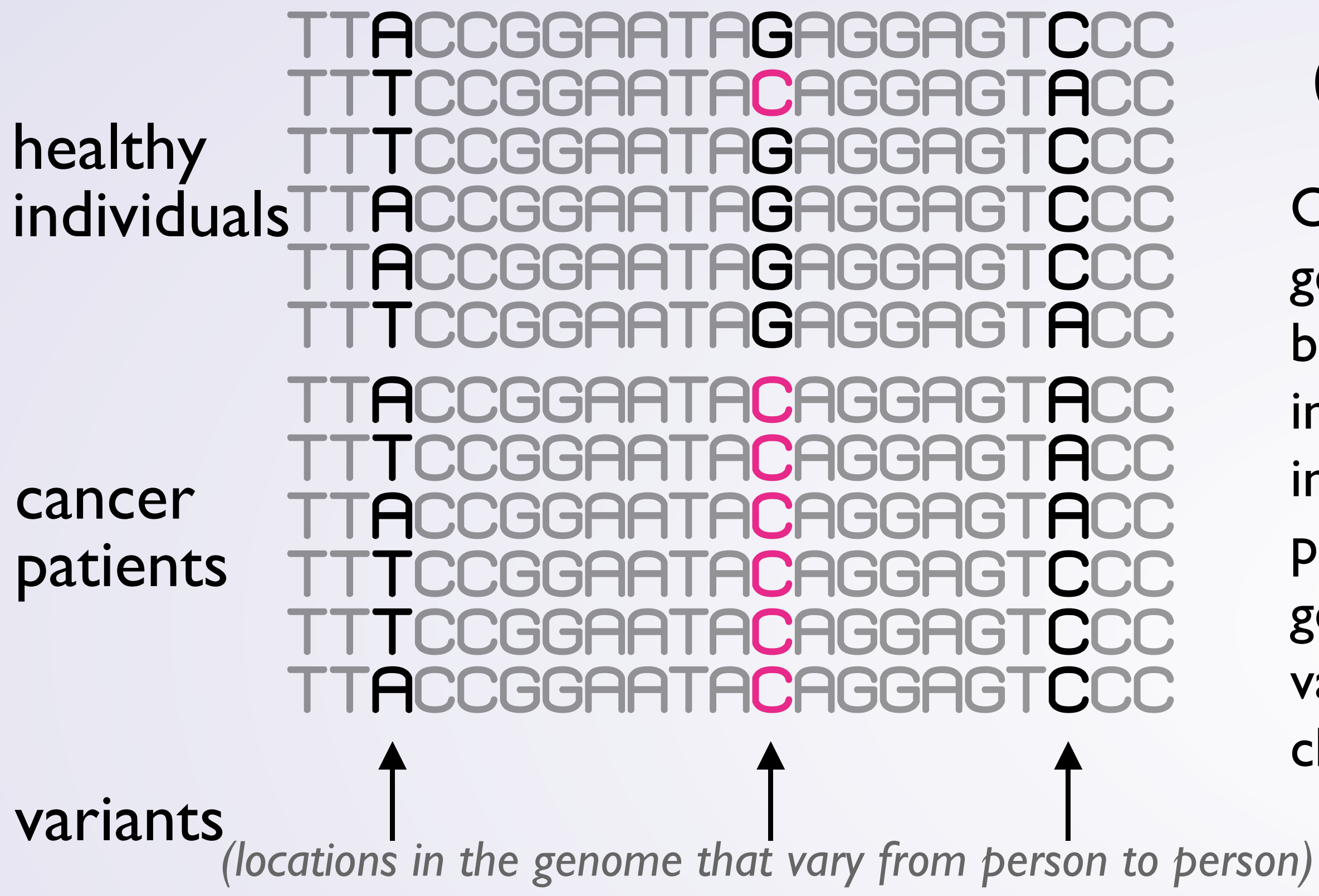
# compreheNGSive

## a tool for exploring next-generation sequencing variants

www.sci.utah.edu

THE UNIVERSITY OF UTAH

alex bigelow
miriah meyer
nicola camp



## data

*what are we dealing with?*

Our collaborators are using next-generation sequencing data to study breast cancer. They are particularly interested in *variants*, or specific locations in the genome that are different from person to person. If we represent the genome as a series of letters, changes at a variant's location will be either a letter change, missing letters, or inserted letters. Specifically, the biologists are looking for sets of variants that change in similar ways across subsets of a population. In this example, the frequency of the letter C in cancer patients is high, indicating that that variant might be relevant to the disease.

## motivation

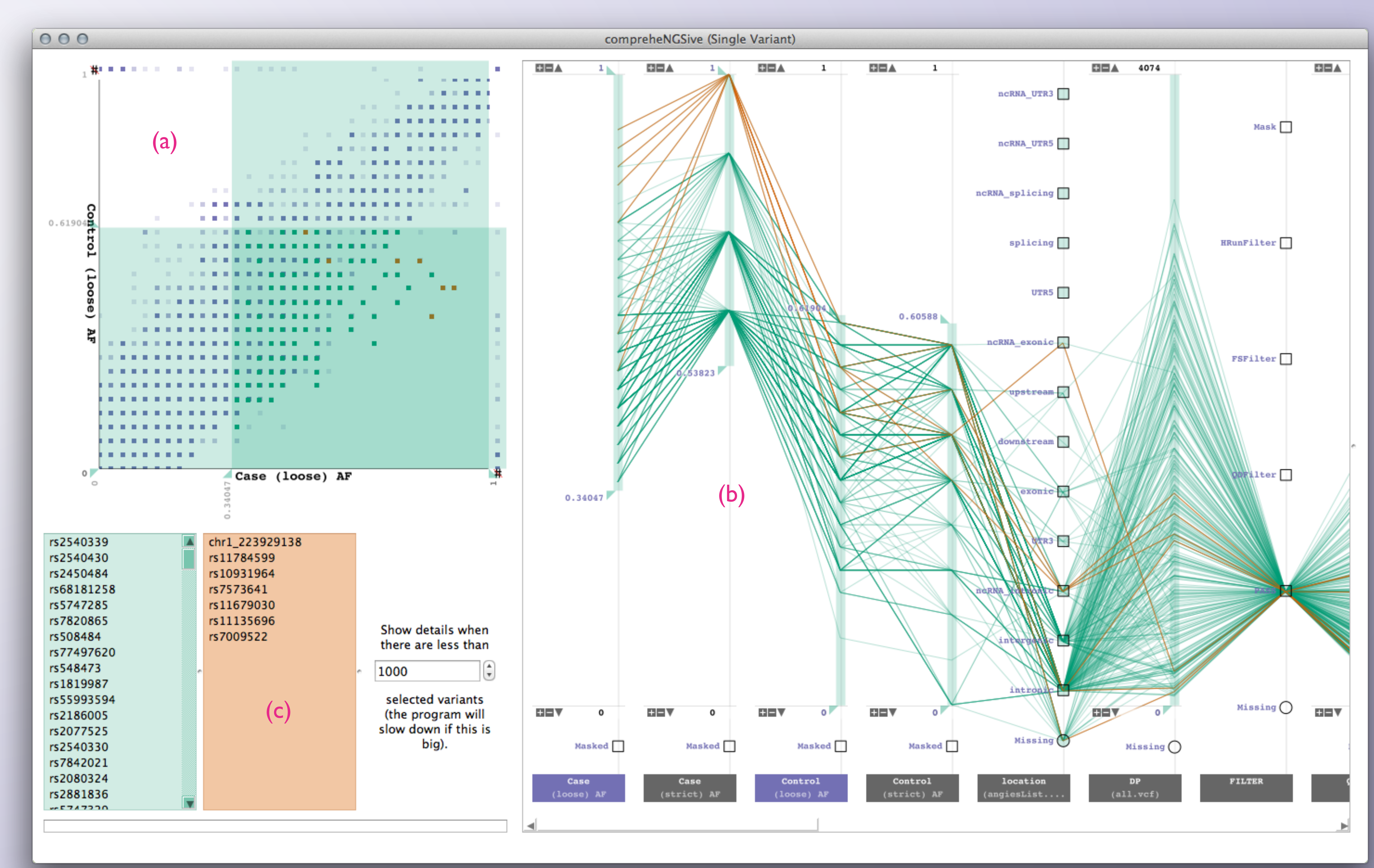
*what are the problems?*

The cost of NGS data is declining sharply, resulting in an unprecedented amount of information for biologists to explore [1]. However, the processes that generate and handle it are extremely complex; a variant collects many attributes before it reaches the point where it can be analyzed. The time and effort to integrate data from these processes can be overwhelming; even a seemingly simple task of assigning individual genomes to meaningful groups becomes difficult. We observed that much of the data is never analyzed because of these difficulties.

## collaboration

*what makes a variant interesting?*

The most challenging aspect of this design study has been articulation of required analysis tasks — a well-known hurdle in collaborative, problem-driven visualization research [2]. At the start of our collaboration the scientists could only articulate that they wanted to explore *interesting* variants. It took over a year of embedded work with these scientists and multiple prototypes to understand the meaning of *interesting*. In particular, our software prototypes caused the scientists rethink their analysis multiple times as they dug into the data for the first time.



rs2540339
rs2540430
rs2450484
rs68181258
rs5747285
rs7820865
rs508484
rs77497620
rs548473
rs1819987
rs55993594
rs2186005
rs2077525
rs2540330
rs7842021
rs2080324
rs2881836

chr1_223929138
rs11784599
rs10931964
rs7573641
rs11679030
rs11135696
rs7009522

Show details when there are less than 1000 selected variants (the program will slow down if this is big).

## tasks

*the workflow we identified*

1) Each individual is assigned to a group or groups
2) Statistics across groups are calculated
3) A subset of variants is filtered, guided by heuristics and the distribution of the data
4) Variants of interest are explored and prioritized across attributes
5) Lists of high-priority variants are created

## tool

*prototype*

The prototype includes three linked views [3] including:
(a) a scatterplot,
(b) a parallel coordinates view, and
(c) a list view.
Each view is linked by a set of selected variants (green) - a variant is in this set if it passes the criteria determined by parallel coordinate and scatterplot sliders. Mousing over a variant in any view highlights it in all of the views (orange). These mechanisms support tasks 3 to 5.

## plans

*coming features*

- Separate preprocessing tool to support tasks 1 and 2
- Basic genome browser
- Brushing mechanism in the scatterplot view
- To better support task 3 in the parallel coordinates view, an integrated histogram view is being considered.

## acknowledgements

*and references*

[1] Kris Wetterstrand. Dna sequencing costs: Data from the nhgri large-scale genome sequencing program.

[2] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *to appear in IEEE Trans. on Visualization and Computer Graphics (Proceedings of InfoVis)*, 2012.

[3] Jonathan C. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Proc. Intl. Conf. on Coordinated and Multiple Views in Exploratory Visualization (CMV)*, pages 61–71. IEEE Computer Society, 2007.

This work is supported in part by a grant from the Avon Foundation for Women