1/15/17

High throughput vs. low throughput

One measures massive amount of sequencing within short length, such as illumina novaseq.

One functions less effectively, elisa, pcr, western blots…

Company culture: cooperation, unlock the power of genome

1/18/2017

On site was split into 4 components, 2 group activities, 1 technical/behavioral interview, and 1 on the spot presentation.

1. Review resume stuff
2. Prepare for behavioral questions
3. Illumina background

Know about each product line---target customer

**Miniseq:** Cost effective, small, start of NGS, small storage

Coverage/depth=number of times nucleotides is sequenced

Coverage per genome=total output generated/total size of the sample sequenced

Coverage per human genome on illumina hiseq 2500=1000 billion base pair/3 b bp=333X

How to choose proper coverage for research lab? How to choose sequencer?

Under support: coverage calculator, how about having integrated site for choosing sequencer? Including coverage, throughput, application (large, small scale: common disease vs. rare disease) Don't have to fill up the form to get the guide or contact sale representative? More efficient?

Novaseq: only for research based? Diff between research based and diagnostic based?

1/19/2017

Illmina competitor: Affymetrix, life technology corporation (acquire by thermos fisher scientific)

Illumina has more focused management team, CEO Jay Flatley

Illumina instruments aren’t perfect—no machine is—but it has the best combination of speed, price, and accuracy in its product lineup. It can look at gene expression, gather data on certain regions of the genome (genotyping), or it can sequence the whole shebang, the complete 6 billion-letter genome signature of an individual. Illumina was also shrewd to make sure that its smaller MiSeq instrument could be used to double-check the accuracy of runs done on its high-end HiSeq instrument, which makes it so researchers don’t need to buy a different company’s tool for validation runs.

How to use MiSeq for double checking?

**Genotyping** is the process of determining differences in the genetic make-up ([genotype](https://en.wikipedia.org/wiki/Genotype)) of an individual by examining the individual's [DNA](https://en.wikipedia.org/wiki/DNA) sequence using [biological assays](https://en.wikipedia.org/wiki/Bioassay) and comparing it to another individual's sequence or a reference sequence.

“You can see Illumina trying to consolidate it into a big vertically integrated system. Like Apple, they basically have control over the whole workflow from beginning to end,”

As the cost to sequence a gene falls and researchers increasingly discover the genetic causes of disease, drugmakers are flocking to create genetically inspired, personalized medicine. As a result, there's been a tidal wave of demand for gene-sequencing machines that allow researchers to peer into and analyze DNA.

Currently, Illumina, which markets high-throughput machines that can sequence an entire genome for less than $1,000, is the leading manufacturer of these machines, controlling an estimated 90% market share. Globally, Illumina boasts an installed base of more than 7,500 machines, including 300 top-of-the-line HiSeq X sequencers. As a result, sales of Illumina's sequencers and the consumables used to run them topped $2.2 billion last year, up 19% from 2014.

Pacific Bioscience—sequel: far smaller and less heavy, deliver longer read data than Illumina and researchers may find that advantage compelling, especially if they're working in clinical research.

Illumina: financial flexibility, broader product line (mature infrastrature)

Given Roche is a global powerhouse and the role of sequencers as a tool for diagnosing illness and determining treatment protocols could be huge, Roche's sequencing business could pose a big threat to Illumina's MiSeqDx machine, which targets the in vitro diagnostics market.

1/22/2017

Research on company’s annual report

Determine how my skill fits the position (understand your resume, elevator intro)

When preparing to write a cover letter, you may find it helpful to think about your Unique Selling Proposition. What is the one thing that makes you unique? What makes you better than other candidates applying for a similar position with this company? What can you offer that no other applicant can? What is the one reason the employer should want to hire you above all other candidates?

Elevator pitch

I’m currently pursuing my master’s degree in Biostatistics at Johns Hopkins University with focus in clinical genomics and epigenomics. I have strong analytical backgrounds with experience of kaggle competition using classification algorithm such as random forest and boosting in R. Additionally, I have received both SAS advanced and base certificate. One of my greatest strength is that I’m multi-tasked, which indeed will help me to fit into the fast-paced culture of illumina.

Find out how to contact the interviewer. If you don’t hear back, you will need to know whom to contact and whether the employer will accept calls to check the status. Example: “I’d like to stay in touch and follow up with you in a week or two to see how the process is going and where I stand. How do you prefer that I communicate with you — email or phone?”

Get business card ask if provide email may I have ur info

for the Personnel Manage or Human Resources Director: -what do employees like best/least about the company? -how much turnover is there? -how much travel is expected? -child care considerations? -what are the chances of being relocated once on the job? -how are raises/promotions determined? -how often are performance reviews given, and how are they conducted? what type of training program does a new employee receive?

for your prospective supervisor: - what are the major responsibilities of the department? - what would I be expected to accomplish within the first six months/year of the job? - what are some of the special projects now ongoing in the department? - how much contact with management is there? - what are some of the long-term goals of the department?

for a prospective colleague: - what do you like best/least about working for the company/department? - what goes on during a typical workday? - do you feel free to express your ideas and concerns? Do most people? - what are the possibilities for personal growth and advancement? - how long have you been here? Does your future here seem secure?

Take the time and effort to clear your mind of school concerns, deadlines, family concerns, and other problems.

1/23/2017

Resume review

Kaggle Competition

Large scale, over 110k observations. So we were required to build model on training in order to predict animal’s outcome for testing data set.

I think the first thing to handle relatively large data sets is data exploration indeed. That’s why people always say when fitting a model, 90% of the time you are doing tedious data cleaning. Fortunately, the data only contained 12 variables so parameter selection was not computationally difficult. We started with a very simple decision tree model (CART) in r to check for the log loss and computation time. And then gradually implemented more complicated tree algorithm such as random forest. Since random forest utilized bootstrapping, which required to sample the entire data set with replacement multiple times, running this algorithm was highly time consuming. As a team, we had to split work in order to be more efficient. I was in charged of data mining and fitting support vector machine. I think svm’s performance depends highly on the distribution of data. In other words, if none of the kernels fit the data well, then svm is probably weaker in modeling than random forest, which was indeed nonparametric.

(in R, ntree=number of bootstrap sample, mtry=number of random select parameter)

(xyboost, use gradient boost, use simple regression tree on data and then regress on residual, use residual for new prediction, repeat.)

1/25/2017

Datafest

The data sets were given by ticket master. We were given 3 data sets and each contained 50-200 variables. So it was a wide data set and variable selection in this case was extremely hard. We first made our choice to focus on pricing optimization in order to narrow down the variable selection process. Cross validation was not used due to computational intensity. So we first brainstormed about some factors that may affect pricing such as population density and performers’ popularity indeed. Of course population density could be easily figure out by merging the geographical location data with US census data. And we quantified the event popularity by time span, which was normalizing the number of days that 90% of the tickets were being sold.

A classifier using knn was established and visualized using these parameters, but how could we say that the popularity of a single event is different among cities thereby having different pricing strategy? We used annova test to test the time span and defined the significance ratio by TukeyHSD test when you have multiple combinations of events. So the ratio was calculated as number of significant events (p value) divided by the total number of combinations of events (which is n!/2!\*(n-2)!). If this ratio is larger than 0.5, over half of the combinations were truly different in time span so we had more confident to use time span among cities to adjust pricing in order to make more profit.