Data Science → from Strangeness to ATGC

- **01/2007**: PhD in Exp. HE Nuclear Physics from John Harris' group (*Multi-strange Baryon Correlations in Heavy Ion Collisions*);
- Post-docs at STAR & ALICE; CHIMERA
- Between post-docs: a year off, a stay-at-home mother
- **2013**: left Heavy Ions (Livermore Labs) in the wake of The great Budget Sequester

Since **2016** --

Production Data Science person (we have no titles; I've had three or four roles in 5 years) at ***Invitae Inc*** -- Clinical (!!) **Genetics** (!!) Testing lab

Essence of the job: Bioinformatics+Data analysis

betty.abelev@aya.yale.edu

Stole this slide from an internal Data Science presentation

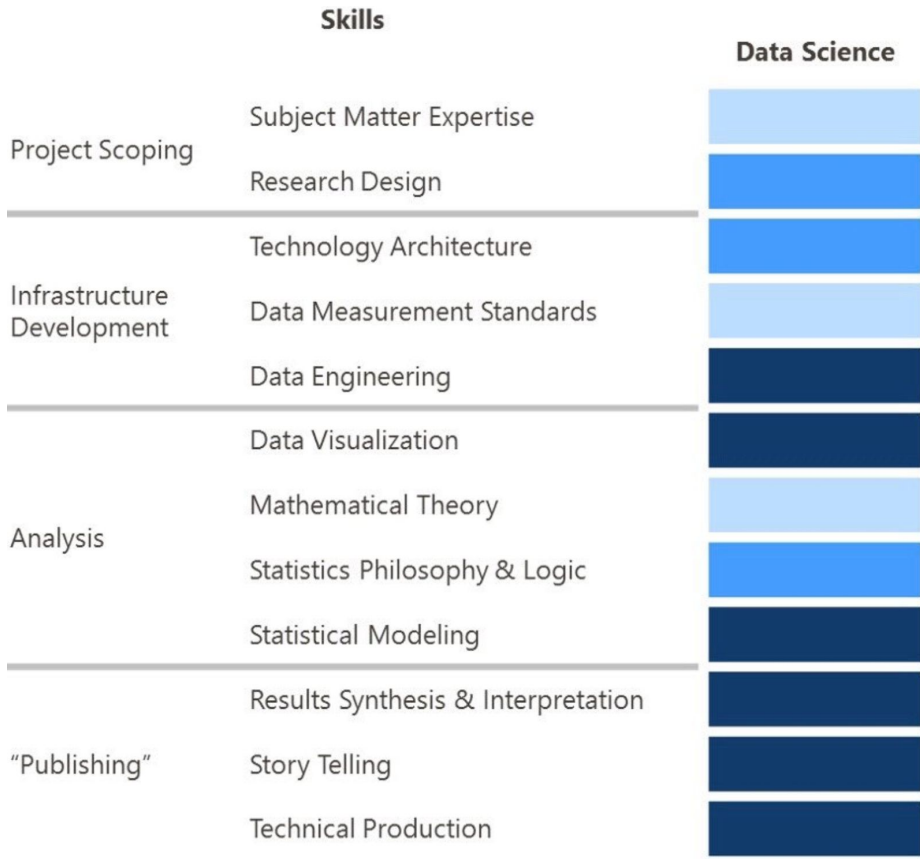
Key skills:

Tenacity (getting to data that's clean)

Knowing what good data looks like

Knowing it's possible and how to tell a story using data

Knowing that fancy modeling methods aren't key

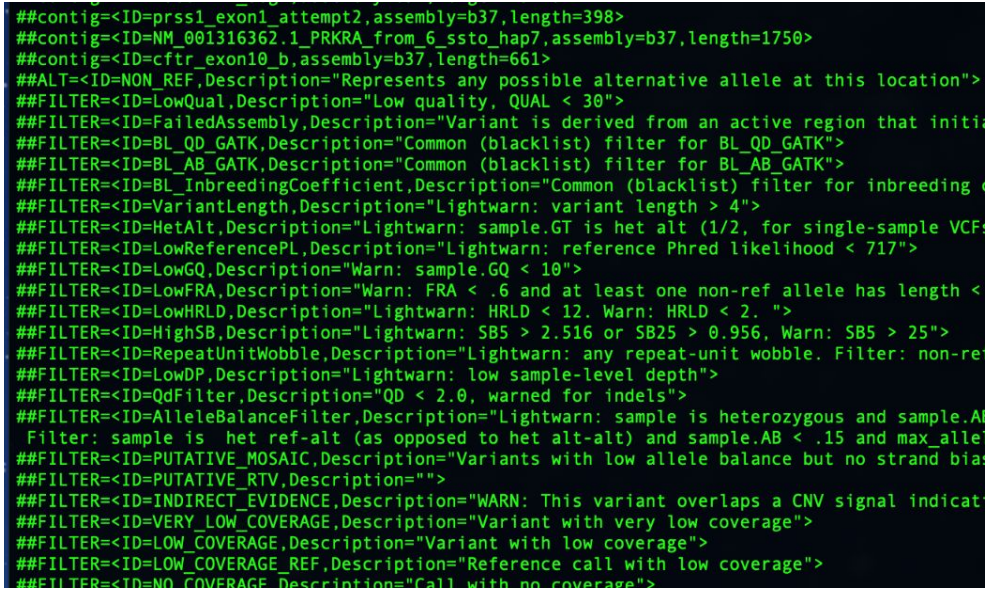


betty.abelev@aya.yale.edu

Biggest culture shock

How genetic variants are delivered -

in multiple file formats
(.bam/.bed/.fastq/.vcf….)


##contig=<ID=prss1_exon1_attempt2,assembly=b37,length=398>
##contig=<ID=NM_001316362.1_PRKRA_from_6_ssto_hap7,assembly=b37,length=1750>
##contig=<ID=cftr_exon10_b,assembly=b37,length=661>
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FILTER=<ID=LowQual,Description="Low quality, QUAL < 30">
##FILTER=<ID=FailedAssembly,Description="Variant is derived from an active region that initi
##FILTER=<ID=BL_QD_GATK,Description="Common (blacklist) filter for BL_QD_GATK">
##FILTER=<ID=BL_AB_GATK,Description="Common (blacklist) filter for BL_AB_GATK">
##FILTER=<ID=BL_InbreedingCoefficient,Description="Common (blacklist) filter for inbreeding
##FILTER=<ID=VariantLength,Description="Lightwarn: variant length > 4">
##FILTER=<ID=HetAlt,Description="Lightwarn: sample.GT is het alt (1/2, for single-sample VCF
##FILTER=<ID=LowReferencePL,Description="Lightwarn: reference Phred likelihood < 717">
##FILTER=<ID=LowGQ,Description="Warn: sample.GQ < 10">
##FILTER=<ID=LowFRA,Description="Warn: FRA < .6 and at least one non-ref allele has length <
##FILTER=<ID=LowHRLD,Description="Lightwarn: HRLD < 12. Warn: HRLD < 2. ">
##FILTER=<ID=HighSB,Description="Lightwarn: SB5 > 2.516 or SB25 > 0.956, Warn: SB5 > 25">
##FILTER=<ID=RepeatUnitWobble,Description="Lightwarn: any repeat-unit wobble. Filter: non-re
##FILTER=<ID=LowDP,Description="Lightwarn: low sample-level depth">
##FILTER=<ID=QdFilter,Description="QD < 2.0, warned for indels">
##FILTER=<ID=AlleleBalanceFilter,Description="Lightwarn: sample is heterozygous and sample.AB
Filter: sample is het ref-alt (as opposed to het alt-alt) and sample.AB < .15 and max_allel
##FILTER=<ID=PUTATIVE_MOSAIC,Description="Variants with low allele balance but no strand bia
##FILTER=<ID=PUTATIVE_RTV,Description="">
##FILTER=<ID=INDIRECT_EVIDENCE,Description="WARN: This variant overlaps a CNV signal indicat
##FILTER=<ID=VERY_LOW_COVERAGE,Description="Variant with very low coverage">
##FILTER=<ID=LOW_COVERAGE,Description="Variant with low coverage">
##FILTER=<ID=LOW_COVERAGE_REF,Description="Reference call with low coverage">
##FILTER=<ID=NO_COVERAGE,Description="Call with no coverage">
1 1957299 . C T 5994.77 PASS ABHet=0.535;DP=435;LIGHTWARN=within_10bp_of_repeat;MLEAC=1;MLEAF=0.5;QD=13.81 GT:GQ:AB:PL:DP:AD:FRA:SAC:SB5:SB25:HRLD:NVAR:FT:GW:LW:ST:DESC 0/1:99:0.465:6023,0,7084:434:232,202:nan:36,196,30,172:0.0:0.0:0.0:13.878:.:PA
1 1959163 . C T 10411.77 PASS ABHet=0.502;DP=657;MLEAC=1;MLEAF=0.5;QD=15.85 GT:GQ:AB:PL:DP:AD:FRA:SAC:SB5:SB25:HRLD:NVAR:FT:GW:LW:ST:DESC 0/1:99:0.498:10440,0,9972:657:330,327:nan:99,231,99,228:0.0:0.0:0.0:15.89:98,1,1:PASS:.:.:CONFIDENT:.
1 1959261 . A G 13174.77 PASS ABHom=1.0;DP=329;MLEAC=2;MLEAF=1.0;QD=40.17 GT:GQ:AB:PL:DP:AD:FRA:SAC:SB5:SB25:HRLD:NVAR:FT:GW:LW:ST:DESC 1/1:99:1.0:13203,986,0:328:0,328:nan:0,0,68,260:0.0:0.0:0.0:40.253:-98,1,1:PASS:.:.:CONFIDENT:.
1 1959978 . C T 3079.77 PASS ABHom=1.0;DP=79;MLEAC=2;MLEAF=1.0;QD=38.98 GT:GQ:AB:PL:DP:AD:FRA:SAC:SB5:SB25:HRLD:NVAR:FT:GW:LW:ST:DESC 1/1:99:1.0:3108,236,0:79:0,79:nan:0,0,1,78:0.0:0.05:0.0:39.342:.:PASS:.:.:CONFIDENT:.
1 1960926 . T C 26032.77 PASS ABHom=1.0;DP=673;MLEAC=2;MLEAF=1.0;QD=38.68 GT:GQ:AB:PL:DP:AD:FRA:SAC:SB5:SB25:HRLD:NVAR:FT:GW:LW:ST:DESC 1/1:99:1.0:26061,2021,0:673:0,673:nan:0,0,461,212:0.0:0.0:0.0:38.724:.:PASS:.:.:CONFIDENT:.
1 1961408 . C T 14504.77 PASS ABHet=0.474;DP=867;MLEAC=1;MLEAF=0.5;QD=16.73 GT:GQ:AB:PL:DP:AD:FRA:SAC:SB5:SB25:HRLD:NVAR:FT:GW:LW:ST:DESC 0/1:99:0.526:14533,0,13200:867:411,456:nan:176,235,221,235:0.08:0.0:16.762:58,1,1:PASS:.:.:CONFIDE
1 1961466 . C T 14487.77 PASS ABHet=0.457;DP=853;MLEAC=1;MLEAF=0.5;QD=17.0 GT:GQ:AB:PL:DP:AD:FRA:SAC:SB5:SB25:HRLD:NVAR:FT:GW:LW:ST:DESC 0/1:99:0.543:14516,0,12380:852:389,463:nan:212,177,249,214:0.0:0.0:0.17.038:-58,1,1:PASS:.:.:CONFIDE
1 2234251 . A G 4453.77 PASS ABHet=0.484;DP=252;LIGHTWARN=not_GIAB_HC;MLEAC=1;MLEAF=0.5;QD=17.67 GT:GQ:AB:PL:DP:AD:FRA:SAC:SB5:SB25:HRLD:NVAR:FT:GW:LW:ST:DESC 0/1:99:0.516:4482,0,4183:252:122,130:nan:120,2,125,5:1.36:0.0:17.786:.:PASS:.:.:LI
1 2236359 . G A 1077.77 PASS ABHet=0.529;DP=68;LIGHTWARN=not_GIAB_HC;MLEAC=1;MLEAF=0.5;QD=15.85 GT:GQ:AB:PL:DP:AD:FRA:SAC:SB5:SB25:HRLD:NVAR:FT:GW:LW:ST:DESC 0/1:99:0.471:1106,0,1267:68:36,32:nan:0.36,0,32:0.0:0.0:0.16.265:.:PASS:.:.:LIGHTWAR
1 2337032 . C T 5865.77 PASS ABHom=0.987;DP=150;LIGHTWARN=not_GIAB_HC;MLEAC=2;MLEAF=1.0;QD=39.11 GT:GQ:AB:PL:DP:AD:FRA:SAC:SB5:SB25:HRLD:NVAR:FT:GW:LW:ST:DESC 1/1:99:0.987:5894,367,0:150:2,148:nan:2,0,148,0:1.38:0.12:39.293:.:PASS:.:.:LIGHTW

.vcf format:

betty.abelev@aya.yale.edu