GATA2 at 14: genotype-phenotype correlations

Amy P. Hsu,[1] Subrata Paul,[2] Jennifer L. Kwan,[3] Philip L.F. Johnson,[4] Justin Lack,[2] Eva P. Szymanski,[1°] Jana Lovell,[1°°] Ladan Foruraghi,[1] Cindy Palmer,[1] Christa S. Zerbe,[1] Janine R. Daub,[1] Joie Davis,[1] Dennis D. Hickstein,[5] Katherine R. Calvo[6] and Steven M. Holland[1]

[1]Laboratory of Clinical Immunology and Microbiology, National Institute of Allergy and Infectious Diseases/National Institutes of Health, Bethesda; [2]Integrated Data Sciences Section, Research Technologies Branch, National Institute of Allergy and Infectious Diseases/National Institutes of Health, Bethesda; [3]Research Technologies Branch, National Institute of Allergy and Infectious Diseases/National Institutes of Health, Bethesda; [4]Department of Biology, University of Maryland – College Park, College Park; [5]Immune Deficiency Cellular Therapy Program, National Cancer Institute/National Institutes of Health, Bethesda and [6]Department of Laboratory Medicine, Clinical Center, National Institutes of Health, Bethesda, MD, USA

°Current affiliation: Internal Medicine, University of Pennsylvania, Philadelphia, PA, USA
°°Current affiliation: Heart and Vascular Institute, The Johns Hopkins School of Medicine, Baltimore, MD, USA

Correspondence: A.P. Hsu
twins@niaid.nih.gov

Received: November 19, 2024.
Accepted: December 29, 2025.
Early view: January 15, 2026.

https://doi.org/10.3324/haematol.2024.287013

©2026 NIH (National Institutes of Health)

Supplementary files to accompany Hsu, et al. GATA2 at 14: Genotype – Phenotype Correlations

Supplemental methods

Data were extracted to a standard form with defined fields for recognized symptoms, age of onset for symptoms, and free-text recording of uncategorized findings. We defined "presenting symptoms" as the first documented GATA2 disease manifestation or atypical illness noted by a physician or recorded in clinical records including lymphedema, hearing impairment, cytopenias, infections, warts, abnormal bone marrow findings, leukemia and non-hematologic malignancy. Lymphedema included reference to lymphedema or documented hydrocele. Hearing impairment included documented congenital deafness, sensorineural or conductive hearing loss or tinnitus. Whenever possible, bone marrow samples were performed or reviewed at our institution by a single hematopathologist, in other cases we relied on outside reports.

Mutations are reported using NM_001145661.2 and NP_001139133.1 as reference genomic and protein sequences respectively. For patients with mutations potentially affecting mRNA processing or stability (premature stop codons or insertions/deletions, and those near splice sites), full-length cDNA analysis was performed when possible to evaluate the effect of the mutation. cDNA amplification of GATA2 was performed as previously reported[1].

Penetrance was defined as the number of mutation positive individuals over the age of 1 year with GATA2 deficiency-compatible clinical symptoms divided by the number of mutation positive individuals. Due to our study selection criteria in which each family has at least one

proband, our penetrance estimates are biased upward relative to a random sample but are directly comparable between mutation categories.

Kaplan-Meier analysis for disease onset and mixed effect Cox regression were performed using R (version 4.2.2)[2] with the survival (version 3.4.0)[3,4] and coxme (version 2.2-18.1)[5] packages. Disease onset was considered to be the age at presence of presenting symptom as defined above and included in Supplemental Table 1. The Cox proportional hazards model is commonly used in survival analysis to model the hazard function which denotes the instantaneous rate of the occurrence of an event at a time given that the subject survived up to that point. Including random effects in the Cox proportional hazards model enables us to incorporate account for familial structure in our data and the resulting non-independence between samples. The kinship matrix (kinship2 R package version 1.9.6)[6] was used to correct for relatedness as a random effect. Code for analysis is available at https://github.com/spaul-genetics/GATA2_Report. Fisher's exact test was calculated for 2 x 2 contingency analyses followed by correction for multiple testing with two-stage step-up method of Benjamini, Krieger and Yekutieli controlling FDR < 0.05.

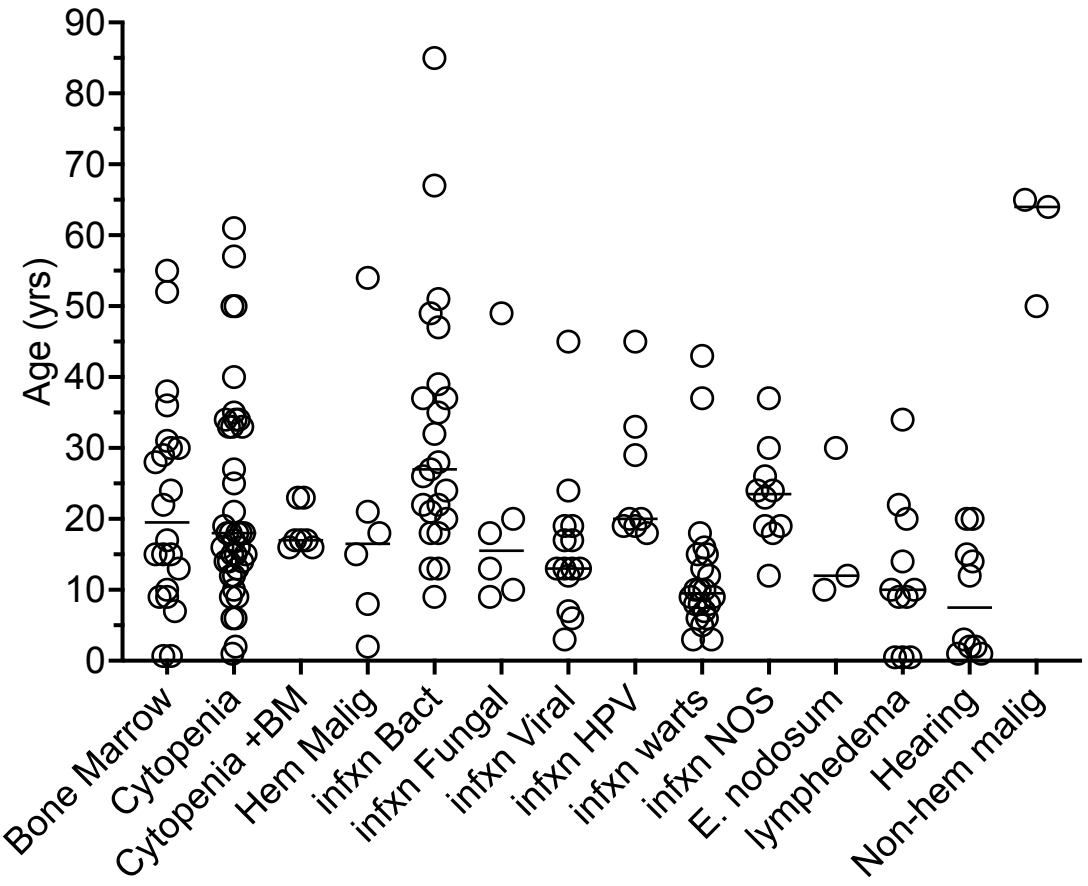
National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) data[7] were obtained from seer.cancer.gov. We investigated the rates of cancers from the SEER 12 registries by age and cancer type for all age groups. Since SEER data is binned in discrete age categories (0-1, 1-4, then 5-year increments), we standardized the GATA2 patients using the age at last patient encounter or transplant. Because GATA2 patients were enrolled over several decades,

and cancer screening processes improved, leading to increased incidence over time, the last year of GATA2 case collection (2020) was selected as the comparator for rates. This choice of operator is expected to bias any estimates downward as an effect of the larger denominator of the most recent data. Since SEER data is patient agnostic, we included each different cancer occurrence among GATA2 patients as a separate event. Data are presented as the proportion of individuals within the age bracket with the specific cancer diagnosis. For leukemia comparisons, “Lymphocytic”, “Myeloid and monocytic”, and “Other” leukemias counts were combined; for female anogenital cancers, “Anus, Anal Canal and Anorectum”, “Rectum”, “Cervix Uteri”, and “Vulva” counts were combined; for male anogenital cancers, “Anus, Anal Canal and Anorectum”, “Rectum”, and “Penis” counts were combined.

Supplementary References

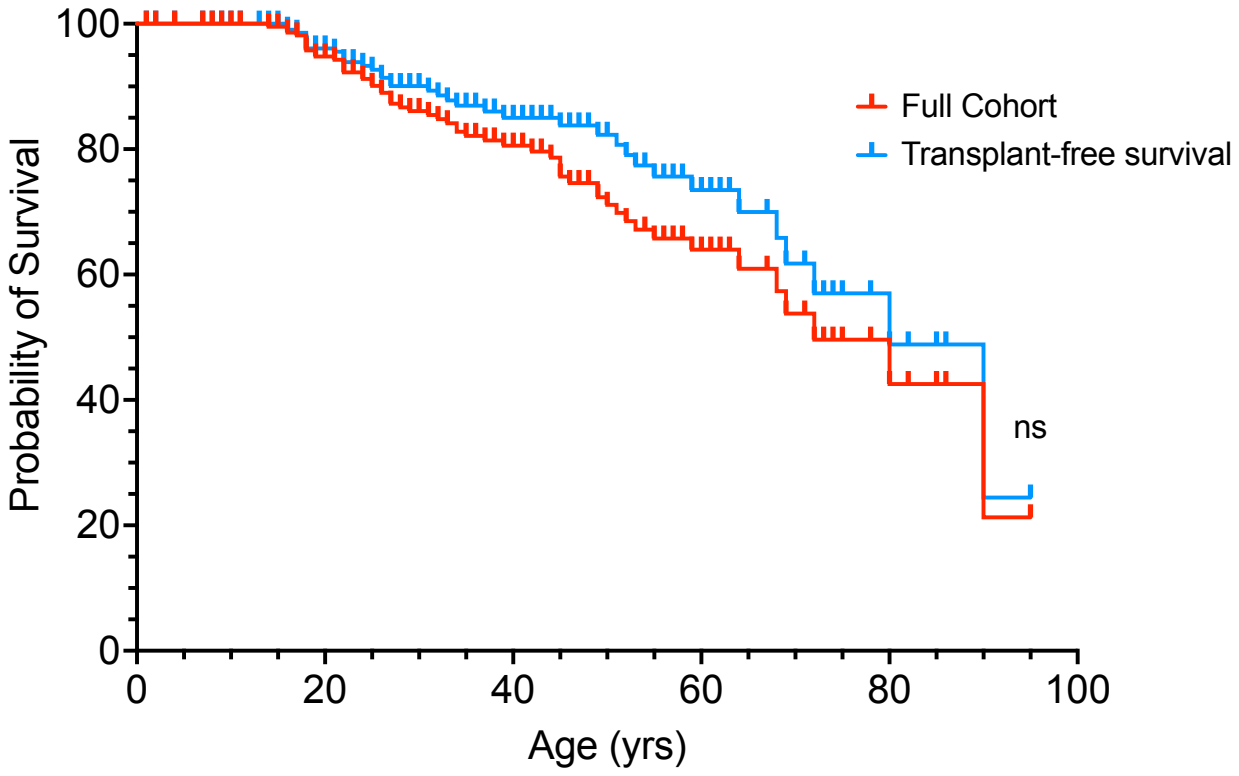
1. Hsu AP, Johnson KD, Falcone EL, et al. GATA2 haploinsufficiency caused by mutations in a conserved intronic element leads to MonoMAC syndrome. Blood. 2013 May 9;121(19):3830-7.
2. R Core Team. R: A language and environment for statistical computing. 2022. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
3. Therneau T. A package for survival analysis in R. R package version 3.4-0. 2022. https://CRAN.R-project.org/package=survival.
4. Therneau TM, Grambsch PM. Modleing survival data : extending the Cox model. 2000. Springer, New York. ISBN 0-387-98784-3.
5. Therneau TM. coxme: Mixed Effects Cox Models. 2022. R package version 2.2-18.1, https://CRAN.R-project.org/package=coxme.
6. Sinnwell J, Therneau T. kinship2: Pedigree Functions. 2022. R package version 1.9.6, https://CRAN.R-project.org/package=kinship2.
7. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER Research Data, 12 Registries, Nov 2024 Sub (1992-2022) – Age-adjusted rates, Linked to county attributes- Time Dependent 1990-2023). National Cancer Institute, DCCPS, Surveillance Research Program, released April 2025, based on the November 2024 submission.

Supplemental Figure 1



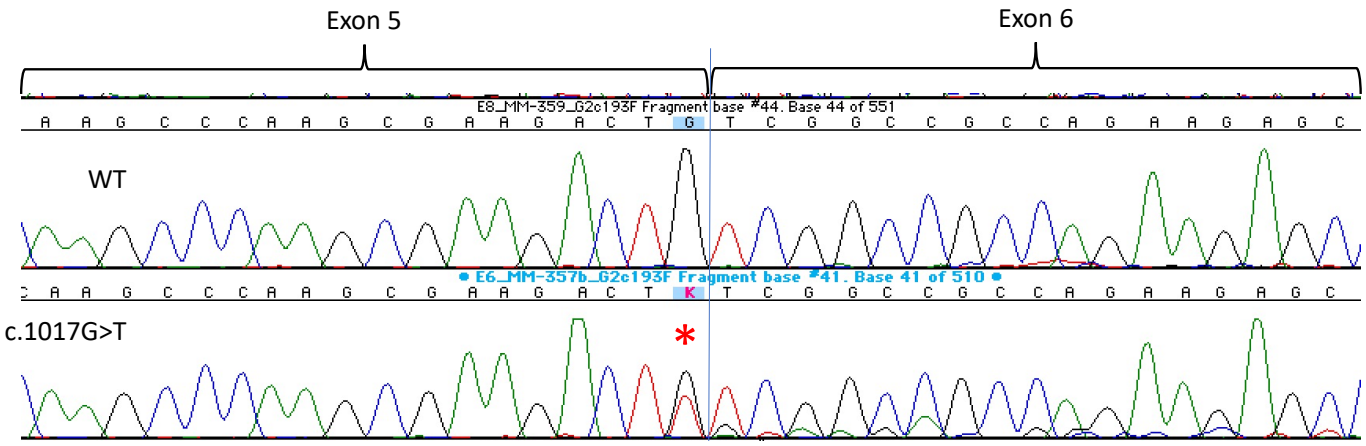
Presenting symptoms across GATA2 cohort. Summary of first symptom observed and age of occurrence.

Supplemental Figure 2.



Kaplan Meier survival curve of GATA2 patients. Red line, survival all patients regardless of transplant status. Blue line, survival censoring patients at age of transplant.

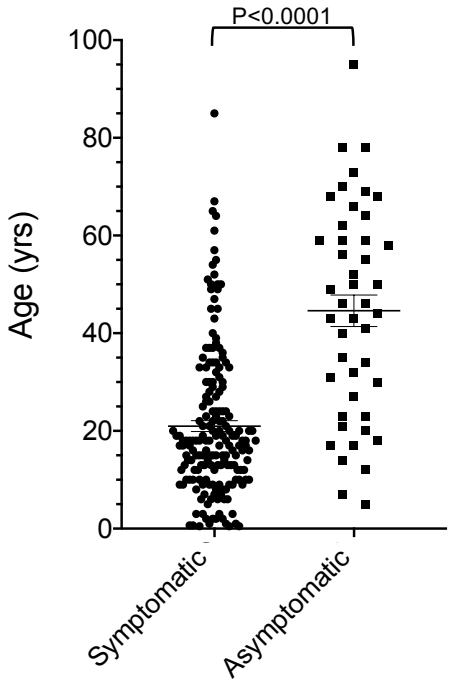
Supplemental Figure 3.



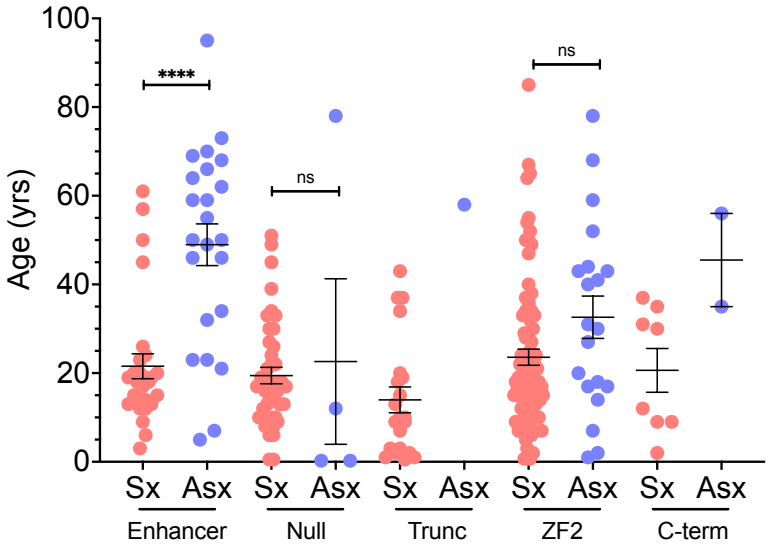
Reduced GATA2 cDNA from c.1017G>T allele. Wild-type sequence shown in top row with appropriate splicing from exon 5 to exon 6. cDNA from c.1017G>T patient peripheral blood mononuclear cells showing the mutation at the last base of exon 5. The mutant transcript is seen in small secondary peaks.

Supplemental Figure 4.

A



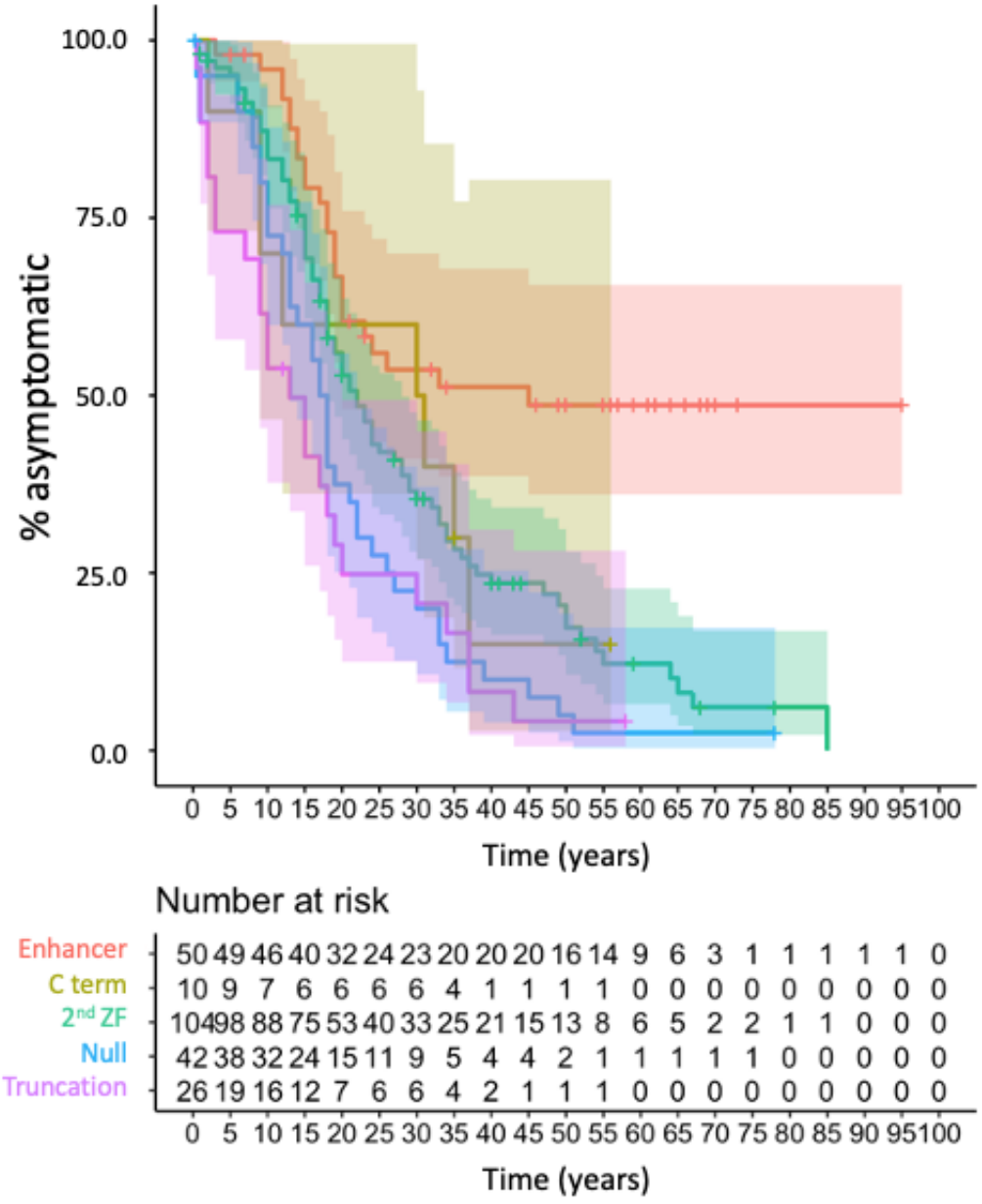
B



Age variation between Symptomatic and Asymptomatic *GATA2* mutation positive individuals.

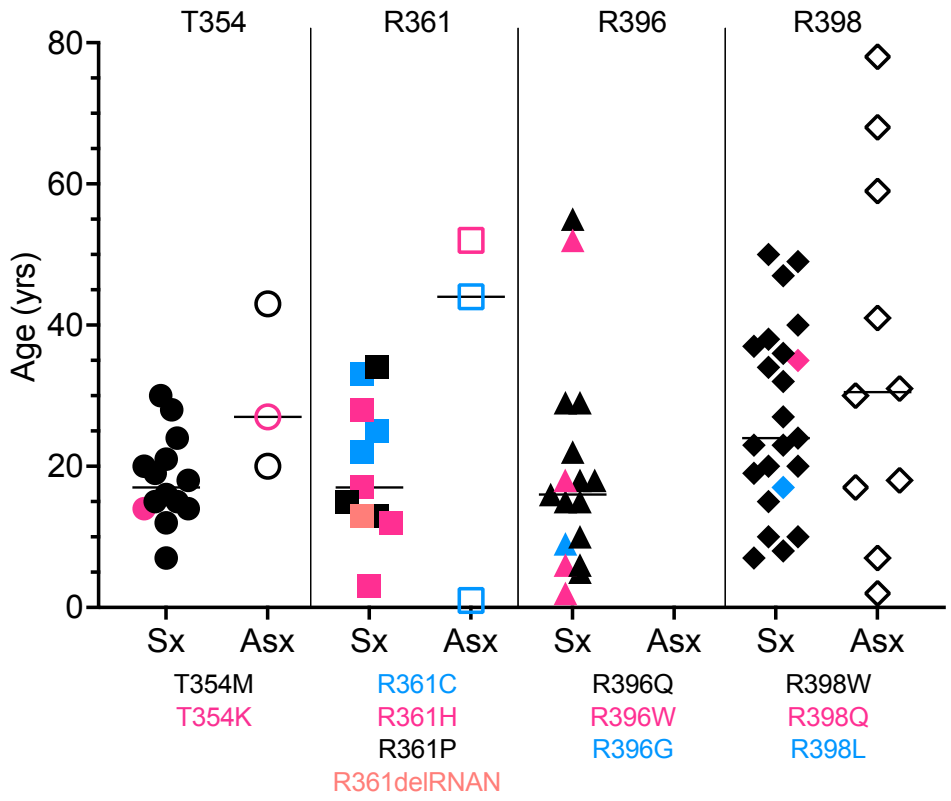
A. Age at symptom onset (Symptomatic) or last clinical visit (Asymptomatic) for GATA2 mutation positive individuals. B. Age at symptom onset as in A with individuals subdivided by mutation group. Bars are median +/- SEM. P<0.0001, Unpaired, two-tailed Mann-Whitney test.

Supplemental Figure 5



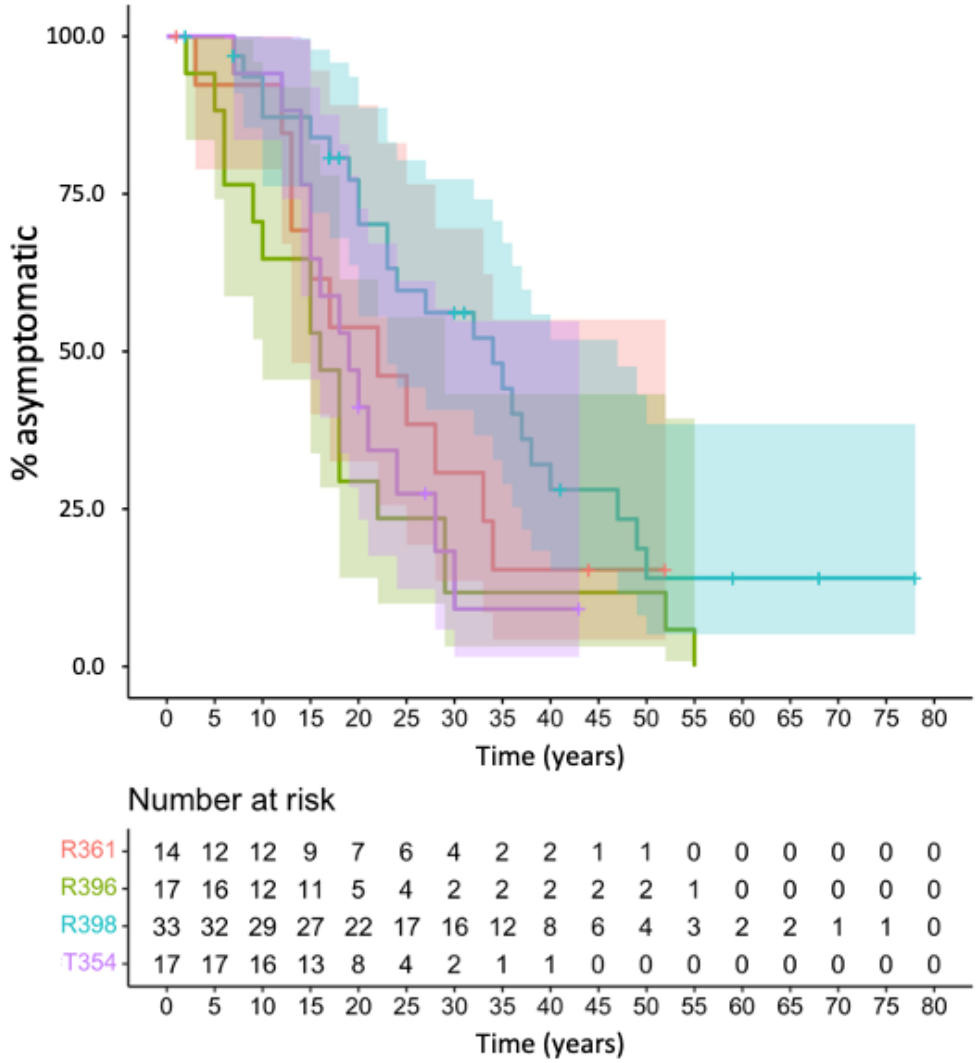
Kaplan-Meier plot of symptom onset across mutation groups. Shading represents 95% confidence intervals. Table below specifies number of at-risk individuals at each 5 year age increment.

Supplemental Figure 6



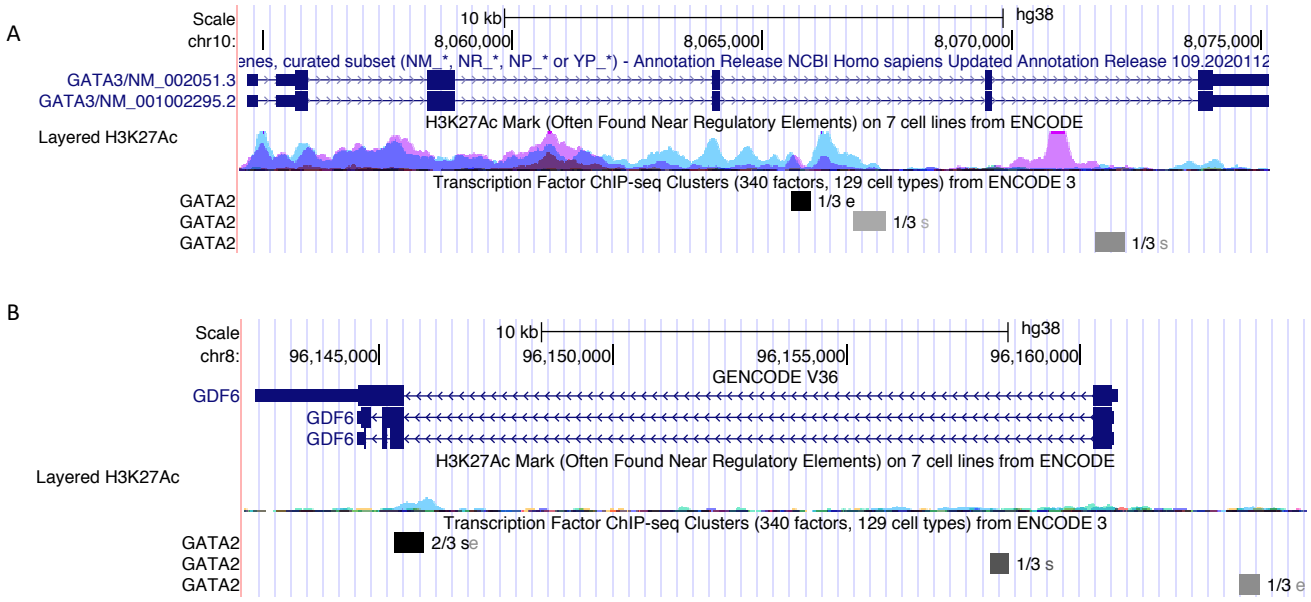
Individuals with mutations at recurrently mutated amino acids. Patients with mutations at one of 4 recurrently mutated amino acids, T354, R361, R396, and R398 are shown, separated by those having symptoms consistent with GATA2 deficiency, (symptomatic – Sx) or remaining asymptomatic (Asx). Age is age of symptom onset for symptomatic patients or age at last evaluation (Asx).

Supplemental Figure 7



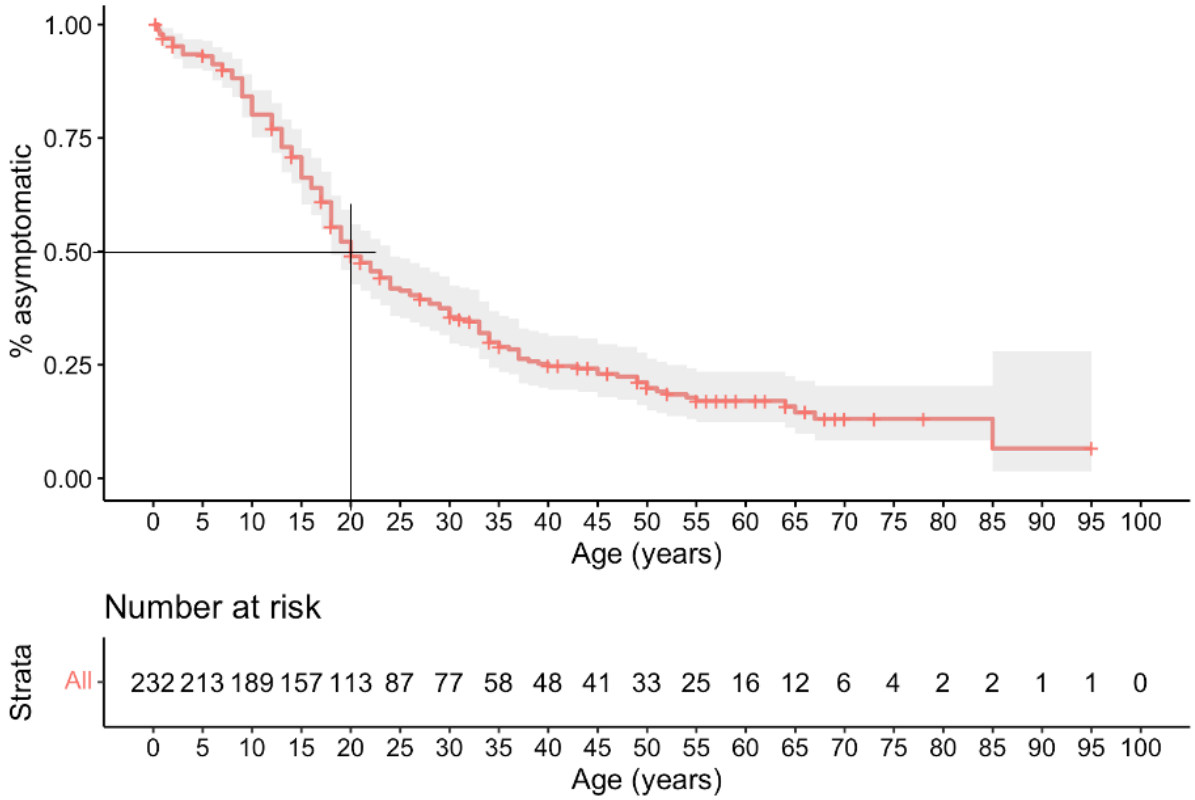
Kaplan-Meier plot of symptom onset by recurrently mutated amino acid. Plot represents individuals with mutations at one of the 4 recurrently mutated, 2nd ZF amino acids, shading indicates 95% confidence interval. Table below specifies number of at-risk individuals at each 5-year age bracket.

Supplemental Figure 8



GATA2 transcriptionally regulates genes critical for inner ear development. A. GATA2 binds to intronic elements within the *GATA3* locus. Shown are two transcripts of GATA3 reflecting exons (boxes) and introns (line with arrows indicating direction of transcription. H3K27 marks are taken from the ENCODE data set (www.encode.org). The binding sites are experimentally determined as part of the ENCODE data set and occur within a region of H3K27 histone acetylation, a marker of regulatory elements. B. GATA2 binds to an intronic element within *GDF6*. Similar to A, there are multiple recognized *GDF6* transcripts, GATA2 binding sites are experimentally determined as part of the ENCODE data set.

Supplemental Figure 9



Median age of symptom onset across entire cohort is (n = 232) is 20 years.

Supplemental Table 1. Data table for all patients included in the study documenting mutation, major symptoms, age at initial symptom onset, bone marrow findings, and current or latest status of patients. This is a separate downloadable file.

Supplemental Table 2. Patient status at last encounter

Status	Number	Age last encounter	Median Age
Alive	104	1 - 95	39.0
Alive / Transplanted	91	10-60	28.00
Deceased	55	14-90	31.00

Supplemental Table 3. Pairwise comparison of mutation categories

Mutation category comparison	**HR**	**p.value (Adjusted)**	**p.value (Unadjusted)**	**Lower CI**	**Upper CI**
Truncation / Enhancer	5.635	**0.000155**	0.0000159	1.889	16.807
Null / Enhancer	3.547	**0.00114**	0.000121	1.445	8.707
Truncation / 2nd ZF	2.52	**0.0487**	0.00619	1.003	6.329
2nd ZF / Enhancer	2.236	**0.0497**	0.00634	1.001	4.998
Null / 2nd ZF	1.586	0.36	0.0684	0.795	3.163
Truncation / C-term	2.531	0.473	0.101	0.539	11.875
C-term / Enhancer	2.226	0.576	0.139	0.509	9.74
Truncation / Null	1.589	0.718	0.209	0.581	4.341
Null / C-term	1.593	0.896	0.367	0.389	6.516
2nd ZF / C-term	1.004	1	0.993	0.259	3.895

Supplemental Table 4. Cytogenetic abnormalities identified by sex

Cytogenetics	Males (n = 92)	Females (n = 111)
-7	11	4
-7, +8	4	1
+8, +1 der(1;7)(q10;p10)	2	1
Chr 7 abnormalities	17	6
+8	4	19
+8, +1(der)	1	0
+8, -11	0	1
Chr 8 abnormalities	5	20
dup 1q	0	1
t(3;18)(q26;q23)	0	1
del 13q	0	1
del Y	1	0

Supplemental Table 5. Frequency of infections and cytopenias

Symptomatic (n = 183)	**Presenting (n; %)**	**Occurrence (n; %)**
Infection	81; 44.3	145; 79.2
Cytopenia	47; 25.7	161; 88.0
Viral		114; 62.3
Genital/ extragenital HPV		94; 51.4
EBV		22; 12.0
Bacterial		74; 40.4
NTM		53; 29.0
staphylococcal		5; 2.7
streptococcal		5; 2.7
Fungal		44; 24.0
Candida		18; 9.8
Aspergillus		12; 6.6

Incidence of infection or cytopenias as presenting symptoms or total occurrence among 183 symptomatic GATA2 patients. HPV – human papilloma virus; EBV – Ebstein Barr virus; NTM – nontuberculous *Mycobacteria*.

Supplemental Table 6. Bone marrow abnormalities and classifications

Bone Marrow (n = 167)	**Occurrence (n; %)**
Any abnormality	142; 85.0
MDS	102; 61.1
In-house review	116; 69.5
WHO MDS	61; 52.6
MDS-MLD	43
MDS-EB1	4
MDS-SLD	4
MDS-MPN	3
MDS-U	2
RCC	2
CMML	2
MDS-EB2	1

Bone marrow findings from clinical records of 167 patients; 116 bone marrows were reviewed by a hematopathologist, 61 patients met the WHO criteria for MDS. Identified subtypes are listed in order of prevalence. MDS – myelodysplastic syndrome; WHO MDS – MDS according to World Health Organization criteria; MDS-MLD – MDS with multilineage dysplasia; MDS-EB1 – MDS with excess blasts; MDS-SLD – MDS with single lineage dysplasia; MDS-MPN – myelodysplastic/myeloproliferative neoplasm; MDS-U – unclassifiable MDS; RCC – refractory cytopenia in childhood; CMML – chronic myelomonocytic leukemia; MDS-EB2 – MDS with excess blasts.